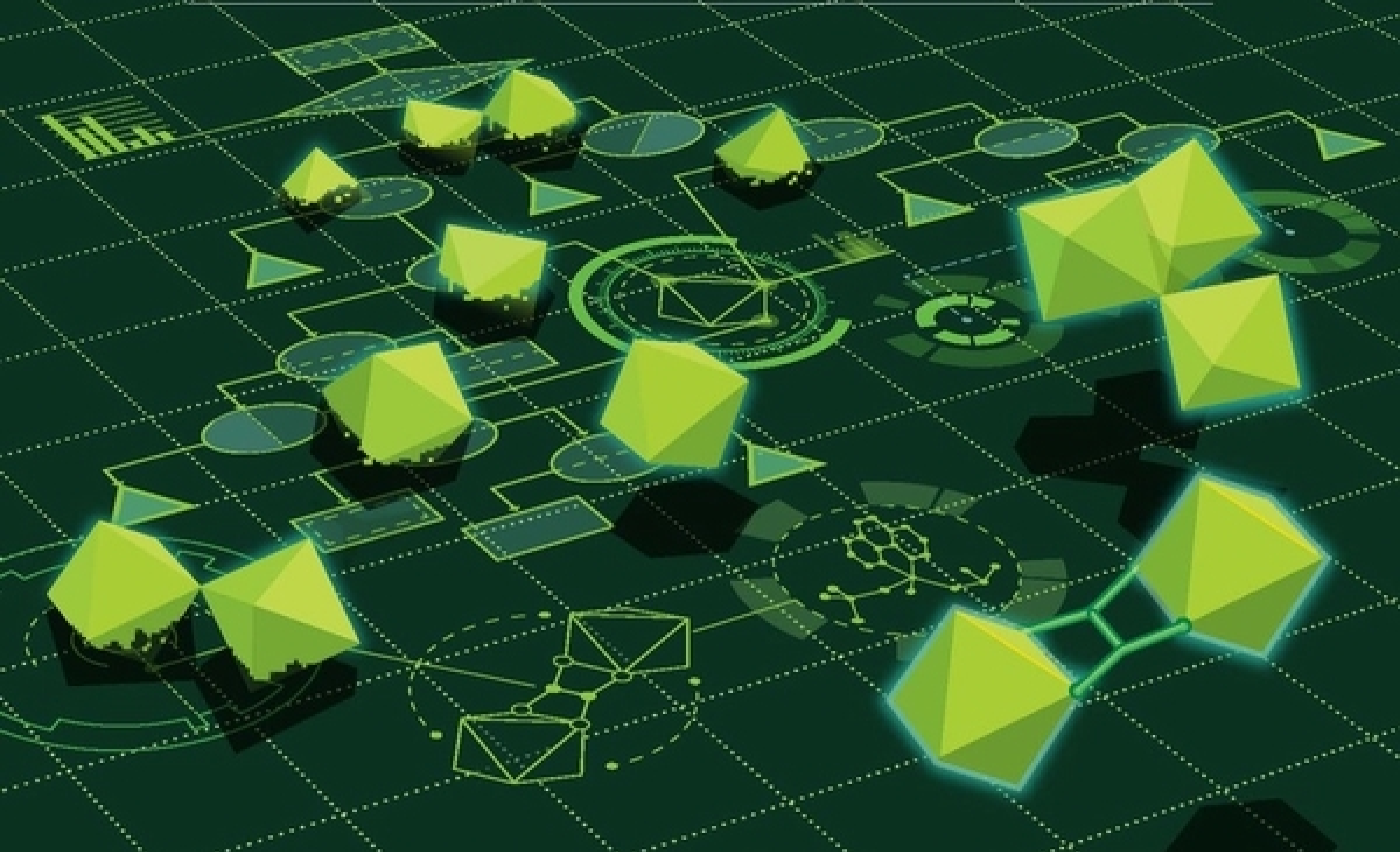
OUTLOOK
*Research commercialization*

# nature

THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE



# LEARNING FROM FAILURE

Machine-learning algorithm mines unreported 'dark' reactions to predict successful syntheses PAGE 73

COMPUTATIONAL SCIENCE

## MATERIALS ADVANCE

*Machine-learning tests all the options — and more*

PAGE 22

PANDEMICS

## STRATEGIC DEFENCE

*New money, coordination and engagement are vital*

PAGE 29

NUCLEAR PHYSICS

## THE BEST OF TIMES

*Elusive $^{229}$Th transition may make nuclear clocks tick*

PAGES 44 & 47

NATURE.COM/NATURE
5 May 2016 £10
Vol. 533, No. 7601

# Lessons from the Ancient One

*The final stages of a dispute over an ancient Native American skeleton signal the need for clearer oversight of such human remains.*

The decades-long battle over the fate of the remains of an 8,500-year-old human known as Kennewick Man may be nearing an end. Last week, the US government determined that the remains are Native American and are thus governed by a law that provides for the repatriation of Native American remains and cultural artefacts.

Five tribes are seeking custody of the bones, and if any can now demonstrate that Kennewick Man is one of their own, they will get the reburial that they have been asking for since the remains were found on the banks of the Columbia River near Kennewick, Washington, in 1996.

The return of the Ancient One, as the tribes call the ancient human, would help to heal a rift between researchers and Native Americans. It also demonstrates the need for a rethink of the rules. In an age in which ancient genomes can reveal startling links between historical populations, we should ask not just whether remains should be reburied, but who decides and on what grounds.

Kennewick Man's genome, reported last year in this journal (M. Rasmussen *et al. Nature* **523,** 455–458; 2015) paved the way for the US Army Corps of Engineers, which manages the land where the remains were found, to deem him Native American. Before that, the bones were in limbo and kept off display, but were allowed to be visited by scientists and the tribes seeking reburial.

The genome established that Kennewick Man is more closely related to Native Americans than to other global populations sampled. This was no surprise and it torpedoed fringe theories that Kennewick Man was related to Europeans or an indigenous Japanese group.

But the researchers also found that some South American groups such as the Karitiana, who live deep in the Amazon, are more related to Kennewick Man than are many North American tribes, such as the Ojibwa from the Great Lakes region. Of the five tribes seeking reburial, only members of the Confederated Tribes of the Colville Reservation offered their DNA for comparison. Members of this tribe were found to share a relatively close connection to Kennewick Man, but no more than some other groups from North and South America.

This ancestry offers a glimpse at the peopling of the Americas, which probably began some 15,000 years ago when groups from Asia crossed the Bering land bridge into what is now Alaska. Researchers are still piecing together this trek, and it is one of the most exciting areas of human population genetics research. Evidence from ancient and contemporary genomes suggests that the journey was far from simple: multiple waves of humans probably settled on the continents, later moving around and replacing earlier inhabitants as they went.

Kennewick Man's genetic relationship to contemporary Native Americans, including the Colville tribes, will factor into the next decision that the US government faces: whether any tribe can make a legitimate claim to his bones. To make a case, tribes will need to establish a cultural affiliation with Kennewick Man on the basis of several lines of evidence including archaeological, geographical and biological links.

This is where things get tricky. Members of the Colville and the other four Washington-state tribes seeking reburial may be descendants of Kennewick Man, but so too may be lots of other groups, including some in South America. Could the Karitiana also claim the remains?

It is possible that researchers could find people more closely related to Kennewick Man than members of the tribes (who share a history of intermarriage and probably have similar connections to Kennewick Man). There are huge gaps in the understanding of Native American genetic diversity. And DNA analysis can reveal unexpected links. A study last year found that the Karitiana and another Amazonian group have an unexpected kinship with Aboriginal Australians (P. Skoglund *et al. Nature* **525,** 104–108; 2015).

***"Genomic analysis is a powerful tool that is redrafting human history."***

Genomic analysis is a powerful tool that is redrafting human history. But the US government should use its broad-brush insights cautiously as it considers the fate of remains.

The Ancient One will probably end up back in the ground, and many scientists will lament the loss. But there are hopeful signs that disputes such as this between researchers and Native Americans will themselves become a relic of the past. A new generation of geneticists is more likely to involve Native Americans in their research, for instance, by drafting plans for the handling of human remains before they are discovered.

Genetics may be equivocal right now on the identity of Kennewick Man's descendants, but such engagement is the best hope to unravel thousands of years of human relationships, to the benefit of all. ■

# The nuclear option

*China is vigorously promoting nuclear energy, but its pursuit of reprocessing is misguided.*

If there's one country that could disprove the old joke among engineers about nuclear power — that nothing can compete with a paper reactor — it may be China. Nuclear power is enjoying a theoretical renaissance in the United States, with researchers advancing a new generation of inherently safe designs and with start-up companies attracting venture capital. But so far, only China has shown the kind of long-term, strategic thinking that would be required to launch a real nuclear revival.

Nuclear engineers from elsewhere know this, and are racking up frequent-flier points on trips to Beijing and Shanghai to support partnerships that may put paper reactors to the test. Already, China is

building a 210-megawatt demonstration of a pebble-bed reactor, led by researchers at Tsinghua University in Beijing. It could come online by next year, marking a first for safer 'generation IV' reactor designs.

The Chinese Academy of Sciences is also working with the US Department of Energy on molten-salt reactors, which were originally developed and tested at Oak Ridge National Laboratory in Tennessee in the 1960s. Researchers at the Massachusetts Institute of Technology in Cambridge are pursuing a partnership to advance an entirely new design that includes elements of both molten-salt and pebble-bed reactors. And the relative newcomer TerraPower, which is based in Bellevue, Washington, and funded by Microsoft co-founder Bill Gates and others, has signed a memorandum of understanding with the China National Nuclear Corporation (CNNC) to pursue the company's 'travelling wave reactor', which is designed to minimize the need for uranium enrichment.

These partnerships illustrate the advantages of international collaboration. China thinks big and moves quickly, and the world may one day reap the benefits. But the country's zeal for advanced nuclear technology has an ominous side: China's latest five-year plan also promotes the reprocessing of nuclear fuel. CNNC officials are currently negotiating with the French nuclear giant Areva to build such a facility.

The promise of nuclear reprocessing has not panned out. The idea dates back to the beginning of the nuclear era, when officials feared a shortage of uranium resources. Plutonium extracted from spent fuel would be redeployed in breeder reactors, which produce more fuel than they consume. But as it turns out, there is more than enough uranium for the foreseeable future. Moreover, the technologies proved expensive, and the risks became all too clear in 1974 when India used reprocessed plutonium in its first nuclear bomb.

For all of these reasons, the United States and many other nations abandoned the idea decades ago. The United Kingdom is closing its reprocessing operations, and the world would be a safer place if countries such as France and Japan followed suit. China should abandon reprocessing before the inevitable bureaucratic momentum builds up. Instead, the country should focus on reducing costs and developing technologies that might enable nuclear energy to play a larger part.

***"China thinks big and moves quickly, and the world may one day reap the benefits."***

As it stands, the short-term outlook is mixed. Some 444 nuclear reactors currently operate around the world, accounting for as much as 11% of global electricity production. Another 64 are under construction, including 22 in China. But many of the existing reactors are getting old and will need to be replaced.

Meanwhile, the public and politicians in many countries are warier than ever after the 2011 Fukushima accident in Japan. An optimistic projection by the International Atomic Energy Agency suggests that global nuclear-power capacity could increase by a factor of 2.5 by 2050. In a pessimistic scenario, the agency suggests that overall nuclear-power production could remain roughly flat.

New reactors have struggled to compete with other forms of energy production, and perhaps the biggest barrier is the huge upfront cost. It is simpler, faster and cheaper, at least in the short run, to build natural-gas-fired power plants, or to install wind turbines and solar systems.

The US Department of Energy is funding nuclear-energy research, with the support of lawmakers on both sides of the aisle in Congress. But what nuclear power really needs is a comprehensive climate policy that puts a price on carbon emissions and rewards all low-carbon energies. Short of that, the nuclear industry's best hope may be China. ■

# Fat lot of good

*Humans' exceptional ability to burn through calories fuels our evolution.*

In an interview last September with *Cyclist* magazine, five-time winner of the Tour de France, Miguel Indurain, was asked about his extraordinarily low heart rate, which story after story had claimed was as low as 28 beats per minute. "Is it true?" the interviewer asked.

"One day we did a medical test and it read 28, so there is some truth in it," Indurain said. "But normally it was a little bit higher." By normally, the cyclist meant that it was usually 30 or 32 beats per minute. And although that have might have been normal for him, it is extraordinary compared with that of the average adult, whose heart bumps along at closer to 60–100 beats per minute.

Indurain is said to have near-super-human heart and lung capacity to go with his glacial pulse. He may also have an unusually low metabolism — a common way to estimate that particular physiological measure is simply to look at the heart rate. The more the heart pumps, the estimate assumes, the faster the body's cells and tissues will be exhausting their reserves. If that is true, then having a slow metabolism would merely confirm that Indurain has a special physiological status. For as a species, humans tend to burn through calories as if they are about to go out of fashion.

We humans are a conundrum to physiologists when it comes to our energy use, because we seem to have evolved an ability to have our cake and eat it, too. Compared with our primate cousins, we breed more and have larger brains — both of which should sap our energy — and yet we live for longer.

This week, biologists offer an explanation. And it is similar to Indurain's answer when he was asked to explain his success on the roads: we simply work harder.

In experiments described online on 4 May, scientists took direct measurements of daily energy use in more than a hundred people and in all other known species of great ape (H. Pontzer *et al. Nature* http://dx.doi.org/10.1038/nature17654; 2016). Chimpanzees, bonobos, gorillas and orangutans all failed to keep up. Every human expended hundreds of kilocalories a day more than any other ape, and the difference is down to greater metabolic activity in our organs.

In other words, humans have evolved to use more energy. We are the original consumer society: our increased demand for physiological energy is driven by our more efficient way of walking, the energy-dense foods such as meat and tubers we have found, and the methods of cooking we have invented and adopted.

The unusually large energy budget of humans presents both an opportunity and a threat. For a start, it helps to power — and to explain the development of — our unusually large and concomitantly energy-hungry brains. We have always been proud of our large brains. Indeed a century or so ago, men of science (and they usually were all men) would routinely measure human heads and weigh their brainy contents to prove our dominance over the beasts. (They did this as well as making false claims on the primacy of certain human groups over others.) But how we found the fuel to maintain such an expensive cognitive prize, where other primates have not, has long been a puzzle.

Then there is the risk. To have a body that needs to be fed more just to exist is a dangerous strategy in lean times, just as use of gas-guzzling motor vehicles is considered antisocial in a resource-constrained world.

The human culture of food sharing helps us to keep the tank filled. So too does what seems to be a uniquely human trait among the primates: the ability to maintain significant fat reserves as a contingency. Even at his slimmest, Indurain would have struggled to match the body-fat content of the average chimpanzee. We may curse its effects today, but human fat tissue seems to have evolved to protect us from ourselves and our unquenchable thirst for energy. It's true: those who struggle to keep those fat reserves under control really can blame their metabolism. ■

**➲ NATURE.COM**
To comment online, click on Editorials at: go.nature.com/xhunqv

# WORLD VIEW *A personal take on events*

# Set up a public registry of competing interests

*The problem of bias in published research must be tackled in a consistent and comprehensive fashion, says* **Adam G. Dunn**.

Before publishing this article, the editors of *Nature* asked me to declare any competing interests. This is routine practice with most journals and is intended to address the serious issue of bias in research. The problem is that after competing interests are disclosed in published research, almost nothing is done with them.

Setting up a public registry of competing interests may provide a way to solve this problem.

Although journals have strengthened their requirements, disclosures are still far from complete. Around half of the studies that involve investigators who hold relevant competing interests fail to declare them. The reasons are rarely the result of a deliberate attempt to mislead readers. Instead, the common causes are inconsistent requirements across journals and negligence.

Some investigators and editors may think that disclosure is a bureaucratic requirement without much practical value. In the current system, it is hard to disagree. There is no reliable guidance on what readers should do when they encounter a competing interest, and no way to know for sure whether competing interests have compromised the integrity of the research findings. Ignoring research that might be biased is clearly wasteful, but allowing it to influence decision-making without knowing whether the results can be trusted might be worse.

Competing interests can cause significant harm by diverting a research consensus away from the truth — from which it can take years to recover. And the complex relationship between the pursuit of knowledge and the pursuit of profit can make such conflicts more likely. For example, internal company e-mails from 2001 from the makers of the diabetes drug Avandia (rosiglitazone) showed the reluctance of the company to publish trial results that may have revealed cardiovascular risk. These risks remained hidden until at least 2007, when an independent meta-analysis was published.

Other competing interests are more subtle. Research undertaken or funded by industry is more easily measured than are ideology, religion, politics or personal relationships, but all of these can influence the design and reporting of research. Defined in this way, competing interests blanket nearly every field of research. There is clear evidence that they are inextricably link to bias. When studies that have competing interests are compared with studies without them, we find consistent differences in how those studies are designed and reported, or whether they are reported at all. Biases are hidden in subtle differences in study design, selective reporting of outcomes, and conclusions that don't match the results. It is difficult even for experts using well-developed tools to identify biases, so how can we expect readers to succeed?

**OUR SYSTEM FOR DISCLOSING COMPETING INTERESTS IS STILL FRAGMENTED, INCONSISTENT AND INACCESSIBLE.**

**➲ NATURE.COM**
Discuss this article online at:
go.nature.com/egzlvq

We need to move beyond occasionally publishing lists of competing interests alongside articles. We need precise, structured and comprehensive reporting of such interests so that we can treat them like any other confounder.

To achieve this, the research community should establish an online database of interests declared by researchers so that we can more precisely determine the association between competing interests and the potential for bias. It should be publicly accessible, available in formats that can be used by humans and machines alike, designed to allow for updates and corrections, and provide a way to uniquely identify researchers. Because of their openness and independence, organizations such as the US National Library of Medicine and the ORCID researcher registry are well placed to act as central locations supporting compliance and standardization. In turn, publishers, funders and institutions can introduce policies that encourage or mandate the use of a registry.

To encourage broad support, it should be easy for journals, institutions, funders and the public to use registry data for their own purposes. For example, a suitable interface could support publishers that want to develop tools to automatically generate disclosure statements by extracting relevant entries.

To judge the risks of bias associated with different forms of competing interests, the registry will need a taxonomy that can consistently map competing interests into a fixed set of classes. These should include employment or funding by companies that may benefit from the research, remuneration paid directly to a researcher, and ideological, religious or political views that may be reasonably perceived to predispose a researcher to reach a certain conclusion.

A comprehensive, accessible record of competing interests could be used to produce more-precise estimates of their impact on research findings. Using these results as a basis, tools could be developed to help readers to interpret individual studies and to flag up uncertainty caused by competing interests to systematic reviewers when they pool the results from multiple studies.

Despite years of improvements by publishers, funders and institutions, our system for disclosing competing interests is still fragmented, inconsistent and inaccessible. Although we can't avoid the fact that people can be swayed if they think they may benefit from distorting their work, we can do much more than to demand complete disclosure and then to do nothing with the information we get back. ■

**Adam G. Dunn** *is a senior research fellow in the Centre for Health Informatics at Macquarie University in Sydney, Australia.*
*e-mail: adam.dunn@mq.edu.au*

# RESEARCH HIGHLIGHTS

*Selections from the scientific literature*

CLIMATE CHANGE

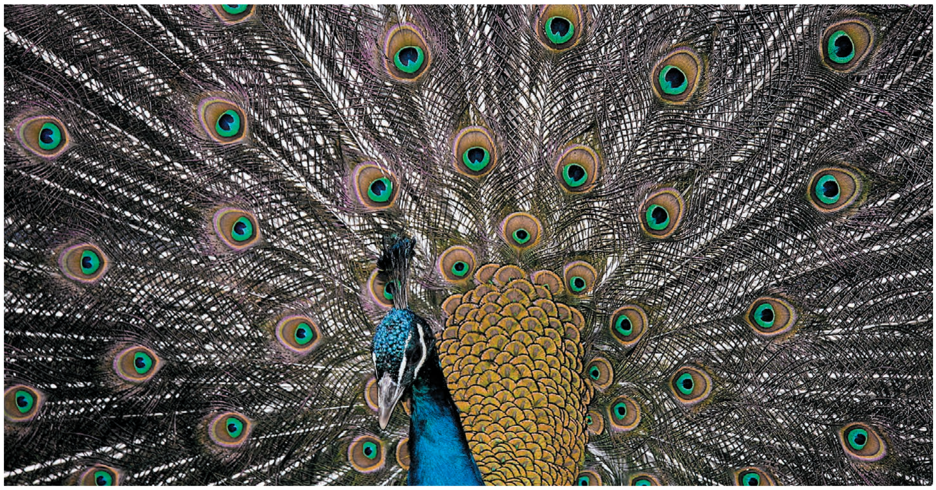
## Knowledge alters public perception

An awareness of the causes of climate change, rather than its consequences or physical characteristics, can increase the public's concern about global warming.

Past studies have suggested that values are more important than knowledge in influencing public perception about climate-change risks. Jing Shi of the Swiss Federal Institute of Technology in Zurich and her colleagues conducted an online survey of roughly 400 people in each of six countries: Canada, China, Germany, Switzerland, the United Kingdom and the United States, controlling for cultural views and values. The results suggest that people are more likely to be concerned about global warming if they understand its basic causes, such as human activities. However, knowledge about the physical aspects of the climate system itself (for example, that burning oil produces carbon dioxide) was correlated with a reduction in concern.

Tailored climate-education programmes might sway public attitudes, the authors say.

***Nature Clim. Change*** **http://dx.doi.org/10.1038/nclimate2997 (2016)**



JUAN GARCIA/GETTY

BIOMECHANICS

## Peacocks maximize tail shimmer

During their elaborate courtship displays, peacocks shake their iridescent tail feathers in an energetically efficient manner.

To study the dynamics of the peacock's tail-feather vibrations, Roslyn Dakin at the University of British Columbia, Vancouver, Canada, Suzanne Amador Kane at Haverford College in Pennsylvania and their co-workers recorded high-speed video of 14 male birds (*Pavo cristatus*; pictured). They found that peacocks rub their tail feathers together at an average frequency of 25.6 hertz, generating a distinct sound and a shimmering effect. In laboratory experiments, the team showed that the feathers resonate when vibrating at this frequency: this maximizes their vibrational amplitude and the shimmering effect.

Scanning electron microscopy also revealed that barbs on the eyespots lock together with microhooks, allowing the eyespots to hold steady.

***PLoS ONE*** 11, **e0152759 (2016)**

IMMUNOLOGY

## Why older people are prone to flu

Minimizing responses from a type of immune cell could help to treat influenza in old people.

The vast majority of influenza deaths are among older people. To find out what makes them vulnerable, Akiko Iwasaki at Yale School of Medicine in New Haven, Connecticut, and her colleagues compared white blood cells from healthy volunteers in their 20s with those from people over 65. When infected with flu virus, cells from the older group produced lower levels of virus-fighting proteins called type 1 interferons.

In mice, knocking out two genes (*Mavs* and *Tlr7*) that help to trigger interferon responses made the animals more vulnerable to both flu and bacterial lung infections. But deleting the *Casp1/11* genes, which help to activate immune cells called neutrophils, protected the mice. The authors suggest that reducing the inflammatory responses of these cells could be a way to treat flu in older adults.

***Science*** 352, **463–466 (2016)**

BEHAVIOURAL ECOLOGY

## Single-celled life can learn

Slime moulds show signs of learning, suggesting that the process does not require nerves and may have evolved early in the history of life.

In a simple form of learning called habituation, an organism learns to ignore continuous stimuli over time. Audrey Dussutour and her team at Toulouse University, France, observed single-celled slime moulds (*Physarum polycephalum*) crossing a bridge in a Petri dish. The bridges were treated with repellent chemicals, either quinine or caffeine, or left untreated. Cells approached and crossed untreated bridges three times faster than cells crossing treated ones. The cells became habituated to treated bridges, crossing them faster after 5 days. However, after 2 days of no chemicals, the organisms' aversion to caffeine or quinine returned. And cells that were habituated to quinine still showed aversion to caffeine, and vice versa, ruling out sensory fatigue or adaptation.

The study suggests that simple learning processes pre-date neuron evolution.

***Proc. R. Soc. B*** 283, **20160446 (2016)**

MICROBIOLOGY

## Gut microbes shape immunity

Autoimmune disease in children could be caused by gut bacteria that inhibit immune development.

Surface lipopolysaccharide (LPS) is made by microbes such as *Escherichia coli* and helps immune cells to mature. Ramnik Xavier of Harvard Medical School in Boston, Massachusetts, and his team studied the gut microbiomes and clinical history of more than 200 children in various countries from birth until age 3. Finnish children, who had higher rates of autoimmune disease than those from Russia, also had higher levels of *Bacteroides* strains than of *E. coli*, whereas Russian children had more *E. coli*. In cultured human white blood cells, LPS produced by *Bacteroides dorei* inhibited the stimulation that is needed to promote immune-system development.

Certain immune-stimulating LPS types might be needed in early life to 'educate' the immune system to more accurately recognize foreign molecules.

***Cell*** **http://dx.doi.org/10.1016/j.cell.2016.04.007 (2016)**

PLANETARY SCIENCE

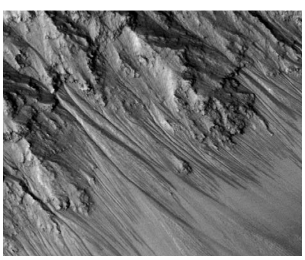
## Martian water on the boil

Water boiling under Mars's thin atmosphere could explain some of the planet's puzzling geological features, such as gullies (**pictured**) and hillside streaks, which some scientists have attributed to liquid water flowing today.

A team led by Marion Massé of the University of Nantes in France melted ice on top of a pile of sand in a laboratory chamber that simulated the Martian atmosphere. Water boiled as it seeped into the sand, causing grains to tumble downhill. Even with relatively small amounts of water, the flowing grains formed channels that were similar to those seen on Mars.

Earth-like quantities of liquid might not be required to form features on Mars, the authors say.

***Nature Geosci.*** **http://dx.doi.org/10.1038/ngeo2706 (2016)**


NASA/JPL-CALTECH/UNIV. ARIZONA

REPRODUCTIVE BIOLOGY

## Catching sperm for contraception

Mouse and human sperm bind to specially designed polymer beads, which could one day be used to select sperm for fertility treatments or to block conception.

Mammalian sperm binds to the ZP2 protein, part of a matrix that surrounds the egg. Jurrien Dean at the National Institute of Diabetes and Digestive and Kidney Diseases in Bethesda, Maryland, and his colleagues attached an engineered portion of ZP2 to the beads, and found that they captured mouse sperm in lab dishes, preventing fertilization of most eggs in the dish. When beads were placed in mouse uteruses, animals gave birth to pups after about 70 days, whereas females with no beads did so after roughly 28 days.

The beads also selected for human sperm in a dish, and, once released, the sperm could bind to and penetrate eggs better than sperm that were not initially captured by the beads.

***Sci. Transl. Med.*** **8, 336ra60 (2016)**

ECOLOGY

## Camera traps may aid conservation

A study using motion-triggered cameras in the wild has revealed that grasslands and floodplains are home to the most diverse communities of mammals in northern Botswana.

Lindsey Rich of Virginia Polytechnic Institute and State University in Blacksburg and her colleagues analysed more than 8,000 photographs of 44 species of mammal taken by 'camera traps' at more than 200 locations across the Okavango Delta of Botswana between February and July 2015 (**pictured** is a female serval; *Leptailurus serval*). They developed models to estimate the spatial distributions of the mammals, and found that species diversity increased with distance into protected areas. Larger species and herbivores benefited from these areas the most, whereas diversity of medium-sized animals was higher in non-protected areas.

The authors say that their methods could be an efficient way of gathering data for conservation of wildlife communities.

***J. Appl. Ecol.*** **http://doi.org/bfqr (2016)**


PANTHERA/LINDSEY RICH

CANCER BIOLOGY

## T cells team up with chemotherapy

Immune cells called T cells could make some chemotherapies more effective against ovarian cancer.

Rebecca Liu and Weiping Zou of the University of Michigan in Ann Arbor and their colleagues studied human ovarian cancer cells in culture. They showed that fibroblasts — connective-tissue cells found in and around tumours — made tumour cells resistant to the platinum-based chemotherapy drug cisplatin by reducing DNA-damaging platinum levels in cancer cells. T cells in the tumour's environment, however, restored the drug's tumour-killing abilities by producing a protein called interferon-γ, which alters certain metabolic pathways in fibroblasts. In women with ovarian cancer, levels of a type of T cell called $CD8^+$ were higher in tumours that were more sensitive to cisplatin.

The results suggest that a combination of platinum-based chemotherapies and drugs that boost T-cell responses could be promising against ovarian cancer.

***Cell*** **http://dx.doi.org/10.1016/j.cell.2016.04.009 (2016)**

ENVIRONMENTAL SCIENCE

## UK food imports use scarce water

Half of the United Kingdom's global water footprint is unsustainable.

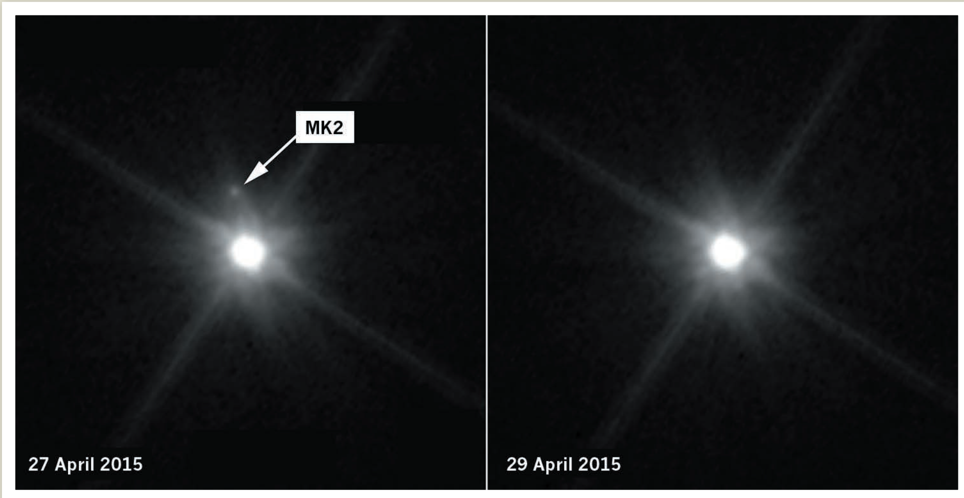
Arjen Hoekstra and Mesfin Mekonnen of the University of Twente in the Netherlands quantified UK water consumption and found that the country uses roughly 5.5 billion cubic metres of surface and groundwater per year. About 5 billion $m^3$ of that is accounted for abroad — mostly water used to produce imported food. Half of this comes from areas that use water unsustainably. For example, almonds are imported from central California, where rivers and groundwater are being depleted to grow this and other crops.

The authors recommend that Britain becomes more self-sufficient in food production and imports food from more water-abundant regions.

***Environ. Res. Lett.*** **11, 055002 (2016)**

**➲ NATURE.COM**
For the latest research published by *Nature* visit:
www.nature.com/latestresearch

# SEVEN DAYS *The news in brief*



NASA/ESA/A. PARKER (SOUTHWEST RES. INST.)

## New moon rises over Makemake

Astronomers have discovered a minute moon around a distant world on the outskirts of the Solar System. Makemake, a 1,400-kilometre-wide dwarf planet in the Kuiper belt, is the second-brightest object, after Pluto, orbiting the Sun beyond Neptune. Its lunar companion, at roughly 175 kilometres across, was nearly hidden in Makemake's glare, but NASA's Hubble Space Telescope spotted it about 21,000 kilometres away from the dwarf planet, astronomers announced on 26 April. The research team, led by Alex Parker of the Southwest Research Institute in Boulder, Colorado, observed the dark moon on only one day in April 2015, so its orbit remains to be established. Dubbed MK 2 for now, the moon will eventually receive a formal name from the International Astronomical Union.

POLICY

### Animals challenge

The European Commission has opened formal proceedings against Italy over how it has adopted the European Union directive on the protection of animals used for scientific purposes into its national law. Italian politicians had added a series of restrictions to prohibit the use of animals in research areas including addiction, and to forbid the use of non-human primates, dogs and cats in basic research. The law will come into effect in January 2017. The Italian government has two months to respond to the commission.

FUNDING

### Cold comfort

Plans unveiled in Australia last week to build a climate research centre in Tasmania offer little solace to many Australian climate scientists who are facing job losses. The Commonwealth Scientific and Industrial Research Organisation (CSIRO) announced on 26 April that it will launch a centre for climate modelling and adaptation in Hobart. The institute, which will have guaranteed funding for a decade, is to employ 40 full-time researchers. But the agency said that it would still be ditching 275 jobs — down from 350 job cuts rumoured earlier this year — in its existing climate, ocean and atmosphere research divisions. See go.nature.com/yvcpcr for more.

EVENTS

### Mission Mars

SpaceX of Hawthorne, California, plans to send an uncrewed spacecraft to Mars as early as 2018, the company announced on 27 April. It is the first time that SpaceX has attached a timeline to its long-standing goal of exploring the red planet. The mission would use a version of SpaceX's Dragon spacecraft — currently used to resupply the International Space Station — but modified with a new propulsion system to descend to the Martian surface. NASA plans to offer technical advice in exchange for data from this 'Red Dragon' mission.

### Native bones

The disputed skeletal remains of a prehistoric person known as Kennewick Man are Native American, the US government decided on 27 April. The US Army Corps of Engineers' decision comes after researchers last year obtained a genome from the 8,500-year-old skeleton and concluded that it was more closely related to present-day Native Americans than to any other population (see *Nature* **522,** 404–405; 2015). The decision paves the way for tribes in Washington state, where the bones were discovered in 1996 near the Columbia River, to seek the reburial of the remains. Until they do, Kennewick Man will remain in storage at a museum in Seattle. See page 7 for more.

### Spaceport debut

A Soyuz rocket blasted off on 28 April to become the first to launch from Russia's newest spaceport, the Vostochny Cosmodrome in Russia's Far East. Vostochny is intended to reduce the country's reliance on the Baikonur Cosmodrome in Kazakhstan, which Russia has leased from the Kazakh government since the dissolution of the Soviet Union. The launch, attended by Russian president Vladimir Putin, carried three satellites into space. Among them was a research mission named Lomonosov, which will study space radiation, including cosmic rays, γ-ray bursts and ionizing events in Earth's upper atmosphere.

### Mars delay

The European Space Agency and its Russian counterpart Roscosmos will shift the launch of their planned ExoMars rover from 2018 to 2020, the organizations announced on 2 May. They blamed the move on delays in European and Russian industrial activities,

as well as late deliveries of scientific payloads. The rover will look for signs of life on Mars, including by drilling as much as 2 metres below the Martian surface.

## Marten shuts LHC

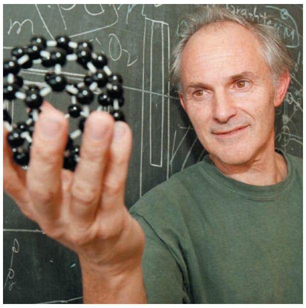
The Large Hadron Collider (LHC) has had its fair share of incidents, but an unlikely one occurred on 29 April, when a beech marten (*Martes foina*) managed to temporarily halt the world's largest particle collider at CERN, Europe's particle-physics lab near Geneva in Switzerland. The animal jumped onto a transformer, creating a short circuit and cutting power to part of the collider. The LHC, unlike the marten, is predicted to recover from the incident.

PEOPLE

## Harry Kroto dies

British chemist Harry Kroto, who shared the 1996 Nobel Prize in Chemistry for the discovery of fullerenes, died on 30 April, aged 76. Fullerenes, elaborate spherical structures of carbon, were discovered in 1985 by Kroto (**pictured**) and colleagues including Robert Curl and Richard Smalley. The researchers named the football-shaped structures after the architect Buckminster Fuller, who designed a dome structure of the same shape. Buckyballs, as the molecules came to be

MICHAEL SCATES/AP/PRESS ASSOCIATION IMAGES



known, are among chemistry's most iconic structures, and are thought to populate interstellar space. Kroto was born Harold Walter Krotoschiner in Wisbech, UK, in 1939 to German parents who had fled the Nazis.

## Airy discord

A prominent French lung specialist could be facing prosecution for having allegedly misled the French Senate about his relationship with the oil industry. Michel Aubier, former head of pneumonology and allergology at the Bichat Hospital in Paris, has been a long-time medical adviser to the French oil company Total. In 2015, he testified on behalf of Paris public hospitals to a Senate commission of enquiry about the economic costs of air pollution. His declaration under oath that he had no ties with "economic actors" could appear misleading, the Senate said in a statement on 28 April. The Senate bureau has asked upper-house president Gérard Larcher to consider filing criminal charges with the public prosecutor against Aubier. Aubier declined to comment.

ENVIRONMENT

## Fungal attack

Asia's first outbreak of a devastating wheat disease is caused by a pathogen that may have arrived from Brazil, a genome analysis released on 26 April suggests. Since February, farmers in Bangladesh have been battling wheat blast, which is caused by the fungus *Magnaporthe oryzae* and has previously been seen only in the Americas (see *Nature* **532**, 421–422; 2016). A team led by Daniel Croll, a microbial population geneticist at the Swiss Federal Institute of Technology in Zurich, found that the Bangladeshi wheat-blast strain is closely related to those circulating in Brazil. Other Asian countries that import wheat from Brazil should watch out for the disease, the team says.

## Iodine provision

Growing concerns over the safety of Belgium's nuclear reactors have prompted the country's government to start supplying iodine pills to its entire population. Health minister Maggie De Block said on 28 April that current precautionary measures, which require pills to be given to residents within 20 kilometres of reactor sites, will be expanded to 100 kilometres, covering all of Belgium. Iodine pills help to prevent the thyroid gland taking up radioactive material during nuclear accidents. Belgium operates seven commercial nuclear reactors; authorities refused to shut down two of them after an independent German reactor-safety commission reported defects in their pressure vessels earlier in the month.

## COMING UP

**6 MAY**
During a close fly-by of the night side of Saturn's moon Titan, NASA's Cassini probe has its only chance to measure the moon's atmosphere while it is receiving minimum external energy.
go.nature.com/ehr6l2

**9–13 MAY**
The European Space Agency hosts the Living Planet Symposium on Earth observation in Prague.
lps16.esa.int

**10–13 MAY**
The fourth international climate-change adaptation conference takes place in Rotterdam, the Netherlands.
go.nature.com/fzdrk2

**10–14 MAY**
At the 29th Biology of Genomes meeting in Cold Spring Harbor, New York, scientists will discuss the role of DNA sequence variation in molecular evolution, population genetics and complex diseases.
go.nature.com/zdewao

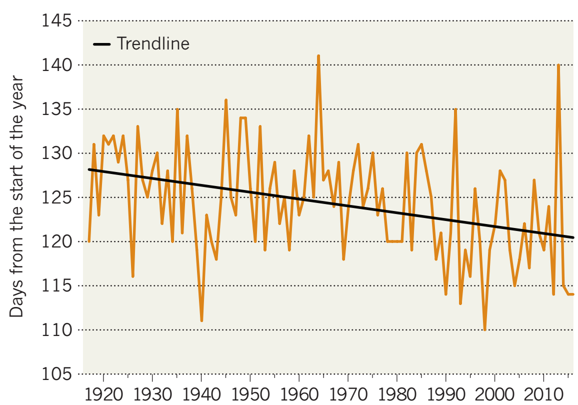
## TREND WATCH

The 2016 Nenana Ice Classic — a lottery to guess when ice in Alaska's Tanana River will break up — ended at 3:39 p.m. Alaska Standard Time on 23 April, after an official tripod lodged in the ice floated 30 metres downstream. In 1917, railroad engineers started betting on when the ice would break up. Records suggest that the spring breakup happens roughly a week earlier than in 1917 owing to climate warming, says an analysis by Gavin Schmidt, director of NASA's Goddard Institute for Space Studies in New York.

SOURCE: US NATL SNOW & ICE DATA CENTER

**BETTING ON THIN ICE**

Data from a century of gambling show that ice on the Tanana River in Alaska is breaking up a week earlier during spring than it did in 1917.



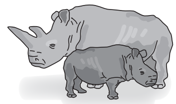

**NATURE.COM**
For daily news updates see:
www.nature.com/news

# NEWS IN FOCUS

**CLIMATE SCIENCE** Australia launches centre in wake of job cuts **p.16**

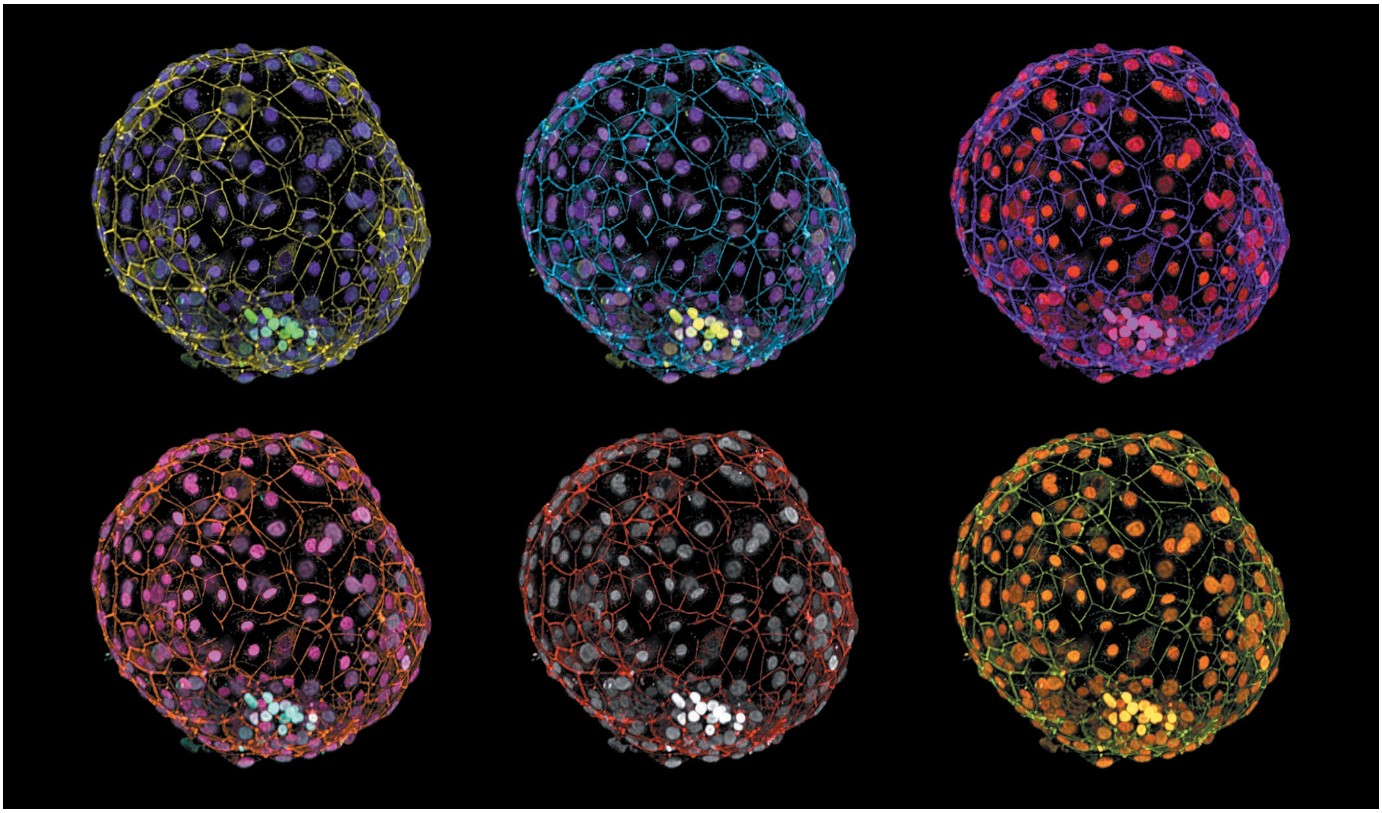
**SPACE** Japanese astronomy satellite doomed by engineering error **p.18**

**CONSERVATION** Stem cells at heart of ambitious rhino rescue plan **p.20**



**MATERIALS** Can machine learning revolutionize materials science? **p.22**

ALESSIA DEGLINCERTI/GIST CROFT/ALI H. BRIVANLOU



**Advances in culturing human embryos (shown here) could reignite ethical debate on the duration of such experiments.**

BIOLOGY

# Human embryos grown in lab for longest time ever

*Embryos cultured for up to 13 days after fertilization open a window into early development.*

**BY SARA REARDON**

Developmental biologists have grown human embryos in the lab for up to 13 days after fertilization, shattering the previous record of 9 days. The achievement has already enabled scientists to discover new aspects of early human development, including features never before seen in a human embryo. And the technique could help to determine why some pregnancies fail.

The work, reported this week in *Nature*[1] and *Nature Cell Biology*[2], also raises the possibility that scientists could soon culture embryos to an even more advanced stage. Doing so would raise ethical, as well as technical, challenges. Many countries and scientific societies ban research on human embryos that are more than 14 days old; in light of this, the authors of the studies ended their experiments before this point.

Scientists have well understood the earliest stages of life in many other animals for decades. "It's really embarrassing at the beginning of the twenty-first century that we know more about fish and mice and frogs than we know about ourselves," says Ali Brivanlou, a developmental biologist at the Rockefeller University in New York City and lead author of the study in *Nature*. "This is a bit difficult to explain to my students."

Magdalena Zernicka-Goetz, a developmental biologist at the University of Cambridge, UK, and her colleagues developed the culture technique using mouse embryos. Many scientists ▶

have attempted to simulate conditions in the womb by growing embryos on a layer of maternal cells, but Zernicka-Goetz's group chose instead to use a gel matrix with higher levels of oxygen. The mouse embryos survived past gastrulation — the stage at which they form layers of cells that will become organs[3]. "It's incredible to look at," Zernicka-Goetz says.

### HUMAN INSIGHT

In *Nature Cell Biology*, she and her colleagues describe how they adapted the technique to work for human embryos donated by an *in vitro* fertilization (IVF) clinic[2]. Zernicka-Goetz and Brivanlou tracked the embryos' progress by comparing the genes that they expressed with those expressed in other animal embryos at similar stages[1]. The scientists were able to evaluate the embryos' structural development using data from a 1956 study in which researchers examined embryos found in women undergoing hysterectomies and other procedures[4].

The teams watched as the cells in the embryos began to differentiate — and reveal features that are unique to human development. For instance, Brivanlou and his colleagues have identified a group of cells that shows up in the embryo around day 10 and disappears around day 12.

The scientists don't yet know the function of the cell cluster, which, at its peak, forms 5–10% of the embryo. But it seems to be a transient organ, akin to the tails that human embryos grow much later in development and then lose before birth. "This is like discovering a new organ in your body," Brivanlou says.

The culture method has also revealed vast differences between the genes expressed in human and mouse embryos, which suggests that rodents may not be good models for understanding human development.

The culture technology is likely to be of broad interest to scientists. Martin Pera, a stem-cell researcher at the University of Melbourne in Australia, says that studying embryos *in vitro* could help researchers who are trying to grow stem cells into embryo-like structures to judge the accuracy of their work.

Once that feat is achieved, scientists could use these structures to conduct larger and more-complicated experiments to explore topics such as the development of birth defects or the effects of toxic compounds.

The fertility industry could also benefit from new *in vitro* technology. Norbert Gleicher, head of the Center for Human Reproduction, an IVF clinic in New York City, notes that about 50% of embryos that implant into a mother's uterus do not survive. Studies of embryos *in vitro* could help researchers to understand what goes wrong in such cases. "The implantation process is a big black box for us clinicians," says Gleicher, who has collaborated with Brivanlou. Gleicher was not involved in the latest work, but he is beginning to use the *in vitro* culture method to study how to evaluate the viability of embryos for implantation in IVF clinics.

> ***"We know more about fish and mice and frogs than we know about ourselves."***

The ability to grow an embryo *in vitro* for 13 days raises ethical and policy considerations. At least 12 countries, including the United States and the United Kingdom, bar scientists from working with embryos older than 14 days. The US government introduced the limit in 1979, on the basis that 14 days marks the beginning of gastrulation in humans. It is also around the latest point at which an embryo can split into identical twins. After this time, the logic goes, a unique individual comes into being.

Zernicka-Goetz and Brivanlou doubt that their embryos would survive much beyond the 14-day mark, because work in mice suggests that more-developed embryos need an unknown mix of hormones and nutrients from the mother to survive. To develop further, the embryos might also require a 3D scaffold to grow on, rather than the flat plates used in the initial tests. To learn more, the researchers are beginning to run experiments with embryos from non-human primates and from cows.

But their achievements in the lab may be grounds for re-examining the limit, says George Daley, a stem-cell researcher at Children's Hospital Boston in Massachusetts. He says that it is somewhat arbitrary. Such a debate would be complex and heated, and it could reach beyond researchers working directly with human embryos. If scientists succeed in growing stem cells into embryo-like structures, it could be difficult to determine whether the structures count as embryos, and thus are subject to the 14-day rule[5]. "It's an interesting ethical discussion we've got ahead of us here," says Pera.

However it plays out, Brivanlou says that the new technology will give developmental biologists plenty to work on. "Every hour as we move forward in development is a treasure box for me," he says. ■ SEE COMMENT GO.NATURE.COM/TQIJ3J


1. Deglincerti, A. *et al. Nature* http://dx.doi.org/10.1038/nature17948 (2016).
2. Shahbazi, M. N. *et al. Nature Cell Biol.* http://dx.doi.org/10.1038/ncb3347 (2016).
3. Bedzhov, I., Leung, C. Y., Bialecka, M. & Zernicka-Goetz, M. *Nature Protoc.* **9**, 2732–2739 (2014).
4. Hertig, A. T., Rock, J. & Adams, E. C. *Am. J. Anat.* **98**, 435–493 (1956).
5. Pera, M. F. *et al. Nature Meth.* **12**, 917–919 (2015).


CLIMATE RESEARCH

# Australian science agency softens blow of climate job cuts

*CSIRO adds 40 posts at new research centre amid hundreds of redundancies.*

BY MYLES GOUGH

After controversially ditching hundreds of jobs in climate research, Australia's national science agency has announced that it will launch a new climate-science centre — but researchers say that the move won't make up for the damage the cuts will cause.

The Commonwealth Scientific and Industrial Research Organisation (CSIRO) said on 26 April that the centre — to be located in Hobart — would employ 40 full-time researchers working on climate modelling, projections and adaptation, and that its funding and staffing levels would be guaranteed for a decade.

But the CSIRO also confirmed details of the job cuts it had announced in February, which have sparked protests in support of Australia's climate scientists. The agency said that 275 jobs would be lost (revising its earlier estimate of 350 redundancies), with about 145 of them in CSIRO's Oceans and Atmosphere, and Land and Water divisions.

"Noting the importance of the climate-science field and following consultation with staff and stakeholders, we determined to maintain a higher level of staffing in this field than flagged earlier in the year," a CSIRO spokesperson told *Nature*.

The new climate centre is "a good news story in terms of what otherwise might have been", says Andy Pitman, director of the Australian Research Council's Centre


JOHN ENGLART/CC BY-SA 2.0

**Scientists have protested against the CSIRO's decision to cut some 300 jobs in climate research.**

of Excellence for Climate System Science in Sydney. "But we don't want to lose sight of the fact that the total scale of capability in CSIRO is being very significantly reduced," he added.

Other scientists were harsher in their judgement. "While the retention of some of CSIRO's climate-science capabilities is welcome, the level announced is analogous to trying to put a sticking plaster over a gaping wound," said Dave Griggs, a sustainability researcher at Monash University in Melbourne, in a statement released through the Australian Science Media Centre.

"This new climate-science centre will be clearly flagging to the international community that CSIRO is committed to a long-term climate-science research capability," Australia's chief scientist, Alan Finkel, told *Nature*. Finkel, who has helped to broker discussions between the CSIRO and climate scientists, acknowledged that there had been "questions raised about CSIRO's reputation" by the cuts.

### CLIMATE PROTESTS

Opposition to the CSIRO's cuts — the result of a strategic shift away from basic climate science — has been strong. Almost 3,000 scientists have signed an open letter to the CSIRO and to Australia's government, raising concerns over the effects of the move on the nation's climate-research capacity. Rallies have been held in major Australian cities, and CSIRO management has been questioned by the Australian senate about its decision, as part of an ongoing inquiry scrutinizing government budget cuts.

But much damage has already been done. One senior scientist from the CSIRO who did not want to be named told *Nature* that senior staff members were already finding new jobs or looking for work elsewhere, and that the organization would find it difficult to keep climate scientists after demonstrating that it does not value their work.

Another researcher — John Church, a specialist in sea-level rise who has worked for the CSIRO for 38 years — says that the new centre is a positive step, but that the overall job losses are "still an incredible cut" to the organization's capability. "You can't hope to cover the range of activities that we did previously when [the CSIRO Oceans and Atmosphere unit] had more than 100 staff, with only 40," he says.

Church says that he expects to be among the scientists made redundant later this year. The reputational damage to the CSIRO is "not going to disappear overnight", he says. ■

## MORE ONLINE

**MORE NEWS**

- Cattle drug threatens thousands of vultures go.nature.com/xrnqce
- UK graphene inquiry reveals commercial struggles go.nature.com/gautwt
- Paper piracy sparks online debate go.nature.com/xpml5h

**NATURE PODCAST**

The value of failed experiments, ketamine without side effects, and our brains' energy demands. nature.com/nature/podcast

STR/AFP/GETTY


The Japan Aerospace Exploration Agency is investigating the factors that led to Hitomi's demise.

ASTRONOMY

# Software error doomed Japanese Hitomi spacecraft

*Space agency declares the astronomy satellite a loss.*

BY ALEXANDRA WITZE

Japan's flagship astronomical satellite Hitomi, which launched successfully on 17 February but tumbled out of control five weeks later, may have been doomed by a basic engineering error. Confused about how it was oriented in space and trying to stop itself from spinning, Hitomi's control system apparently commanded a thruster jet to fire in the wrong direction — accelerating, rather than slowing, the craft's rotation.

On 28 April, the Japan Aerospace Exploration Agency (JAXA) declared the satellite, on which it had spent ¥31 billion (US$286 million), lost. At least ten pieces — including both solar-array paddles that had provided electrical power — broke off the satellite's main body.

Hitomi had been seen as the future of X-ray astronomy. "It's a scientific tragedy," says Richard Mushotzky, an astronomer at the University of Maryland in College Park.

The satellite managed to make one crucial astronomical observation before the accident, capturing gas motions in a galaxy cluster in the constellation Perseus. The instrument that made the observation, a high-resolution spectrometer, had been in the works for three decades. Two earlier versions of it were lost in previous spacecraft failures.

Hitomi's troubles began in the weeks after launch with its 'star tracker' system, which is one of several systems on board designed to keep the satellite oriented in space. The star tracker experienced glitches whenever it passed over the eastern coast of South America, through a region known as the South Atlantic Anomaly. Here, the belts of radiation that envelop Earth dip relatively low in the atmosphere, exposing satellites to extra doses of energetic particles.

> *"We had three days. We'd hoped for ten years."*

By itself, that should not have been a fatal problem. But the star-tracker issue kicked off a series of cascading failures.

At 3:01 a.m. Japan time on 26 March, the spacecraft began a preprogrammed manoeuvre to swivel from looking at the Crab Nebula to the galaxy Markarian 205. Somewhere along the way, the problems with the star tracker caused Hitomi to rely instead on another method, a set of gyroscopes, to calculate its orientation in space. But those gyroscopes were reporting, erroneously, that the spacecraft was rotating at a rate of about 20 degrees each hour. Tiny motors known as reaction wheels began to turn to counteract the supposed rotation.

## SPIN CYCLE

Once the reaction wheels reached their maximum spin, a magnetic rod would normally deploy to keep them from accelerating out of control. But the magnetic rod must be oriented properly in three dimensions to work, and so it failed to slow the reaction wheels. Hitomi spun faster and faster.

The spacecraft then automatically switched into a safe mode and, at about 4:10 a.m., fired thrusters to try to stop the rotation. But because the wrong command had been

uploaded, the firing caused the spacecraft to accelerate further. (The improper command had been uploaded to the satellite weeks earlier without proper testing; JAXA says that it is investigating what happened.)

All of this took place when Hitomi was on the other side of Earth from Japan and unable to communicate with its controllers in real time. In the United States, team scientists went to bed on Friday 25 March, having celebrated what looked like a successful start to the mission. Saturday morning, they woke up to a terse e-mail from the project manager, Tadayuki Takahashi, saying that the spacecraft had been in an emergency.

Ground-based telescopes have since taken pictures of Hitomi spinning roughly once every 5.2 seconds.

### LOST OPPORTUNITIES

Dan McCammon, an astronomer at the University of Wisconsin–Madison, helped to design and build Hitomi's premiere scientific instrument, an X-ray calorimeter that measures the energy of X-ray photons with exquisite precision. He has been working on the technology for more than three decades, flying versions of it on the ASTRO-E mission, which failed on launch in 2000, and the Suzaku spacecraft, in which a helium leak rendered the instrument useless weeks after its 2005 launch.

McCammon says that it would take about US$50 million from NASA, and another 3–5 years, to build a replacement calorimeter. A version of it is slated to fly on the European Space Agency's Athena mission, but that is not due to launch until 2028.

The calorimeter is the biggest loss, says Makoto Tashiro, an astrophysicist at Saitama University in Japan. It was to have gathered extraordinary detail on exploded stars, galaxy clusters, the gas between the galaxies and more. "We lose the new science," he says.

But Hitomi could still contribute to science. Because of the early failure with Suzaku, Hitomi scientists planned one important early observation. About 8 days after launch, Hitomi turned its X-ray gaze on the Perseus cluster, about 250 million light years (77 million parsecs) from Earth. By measuring the speed of gas flowing from the cluster, Hitomi can reveal how the mass of galaxy clusters changes over time as stars are born and die — a test of a crucial cosmological parameter known as dark energy.

That one observation may yield a set of Hitomi papers, says Mushotzky. But no more.

"We had three days," he says. "We'd hoped for ten years." ■

PHYSICS

# Space–time mission draws global interest

*But regulatory hurdles might complicate partnerships in the space-based search for gravitational waves.*

**BY ELIZABETH GIBNEY**

In the wake of the historic detection of gravitational waves by a terrestrial US experiment, a space-borne European effort is drawing interest from a range of parties. But although advisers to the European Space Agency (ESA) recommended increasing international contributions to the billion-euro gravitational-wave detector on 12 April, regulatory hurdles may hinder proposed partnerships with the United States and China.

In February, researchers working on the US-based Advanced Laser Interferometer Gravitational-Wave Observatory (LIGO) announced that they had detected ripples in space-time that had been produced by the merger of two black holes. The space-based observatory planned by ESA would be able to detect ripples with much lower frequencies than would be possible on Earth, bringing into view a greater variety of astronomical events, including mergers between supermassive black holes.

Such a detector is widely seen as "the best thing you could do in gravitational waves", says Robin Stebbins, an astrophysicist at NASA's Goddard Space Flight Center in Greenbelt, Maryland. After a mission to test crucial technologies for the observatory proved successful, the ESA advisory team last month concluded that not only are the agency's plans feasible, but also that the launch could even be brought forward, from 2034 to 2029.

Initially, NASA and ESA were partners in the effort, but funding issues led NASA to pull out in 2011. The US space agency has since stated that it wants only a minor role in the observatory. But excitement around the LIGO findings mean that US scientists are keen for NASA to become an equal partner again, says Rainer Weiss, a physicist at the Massachusetts Institute of Technology in Cambridge who was instrumental in creating LIGO.

Stebbins expects that the committee tasked with assessing progress on the US decadal review, which decides the priorities of NASA and other funding agencies, will express support for a larger role in the ESA observatory later this month. But such a role might require NASA to find more money before the next review, in 2020, and that would mean either diverting money away from other projects or persuading the US Congress to give it more.

Any plan to cooperate with ESA on an equal footing could also come up against ESA's policy of capping international contributions to large missions at 20% to stop projects from falling apart if a partner pulls out. It is too early in discussions to know whether the policy will present a problem, says Fabio Favata, head of science planning and community coordination at ESA.

The United States is not the only country seeking to capitalize on the LIGO breakthrough. Japan's gravitational-wave community is also looking for a way to contribute to the ESA mission. And Chinese scientists have expressed interest for several years now, says Stebbins. They could provide financial or in-kind contributions to the ESA mission in exchange for technical know-how, he says.

US participation could also complicate any potential collaboration between ESA and China. An amendment to US law introduced in 2011 blocks NASA scientists from working directly with Chinese counterparts under almost all circumstances. Stebbins's superiors have told him that the law applies to bilateral collaboration, so it might not apply to a collaboration with ESA that also includes China.

But Congress might try to prevent this kind of collaboration anyway, says Brian Weeden, the technical adviser for the Secure World Foundation in Washington DC, which promotes the peaceful use of outer space. And Congress's scepticism of collaboration with China could stop NASA scientists from even trying to participate. That the gravitational-wave detector is purely a science mission may reassure Congress, Weeden adds. "There may be less concern over that type of cooperation than there would be on cooperation with a more political component, such as human spaceflight."

China is a growing space power — it is scheduled to launch several high-profile space-science missions this year — so the United States will eventually work with China in some capacity, Weeden says. And that would probably be through some kind of multilateral project, he thinks. "The challenge is finding a topic that both the United States and China want to work on. I think the gravitational-wave detector could be one of those." ■

BEN CURTIS/AP PHOTO

Fatu is one of three remaining northern white rhinos, all of which live in Kenya's Ol Pejeta Conservancy.

CONSERVATION

# Geneticists aim to save rare rhino

*Critics say costly plan will divert funds from broader efforts.*

BY EWEN CALLAWAY

The northern white rhinoceros is a species waiting for extinction. Its three remaining individuals, kept in a well-guarded Kenyan conservation park, cannot breed naturally. A 15-year-old female named Fatu could be the last of a creature that once roamed central African savannahs by the thousands.

In a last-gasp effort to avert that scenario, researchers this week unveiled the details of an audacious plan to save the northern white rhino (*Ceratotherium simum cottoni*), by transforming cells from living rhinos and from frozen storage into sperm and egg cells, and then using *in vitro* fertilization (IVF) to create embryos and revitalize the population. Teams led by San Diego Zoo Global in California and the Leibniz Institute for Zoo and Wildlife Research in Berlin have already started work on the idea. They say that it could guide the rescue of other animals that are on the brink of extinction, and even the resurrection of those already gone. But critics call the plan, which is likely to require millions of dollars, fanciful and worry that it could distract from broader conservation efforts.

"The northern white rhinoceros will go extinct if we don't do this," says Oliver Ryder, a conservation geneticist at San Diego Zoo Global and a leading architect of the rescue plan, published on 3 May in *Zoo Biology* (J. Saragusty *et al. Zoo Biol.* http://dx.doi.org/10.1002/zoo.21284; 2016). The strategy was drawn up last December in Vienna at a meeting that was attended by teams from both zoos, as well as specialists in stem-cell and reproductive biology. "It's really a strategic road map — one which has a lot of obstacles," says reproductive biologist Thomas Hildebrandt, who leads the Leibniz team.

## ON THE BRINK

Poaching has slashed the rhinos' numbers from around 2,300 in the 1960s. For the remaining three animals, natural reproduction is not an option. Sudan, a 42-year-old male, has a low sperm count; his 26-year-old daughter Najin has leg injuries that mean she cannot bear the weight either of a mounting male or of pregnancy; and her daughter Fatu has a uterine disorder that would prevent an embryo from implanting. But sperm and other cells from another ten individuals are in frozen storage.

To begin with, researchers will try to create embryos from existing sperm and egg cells; Hildebrandt says that this year he will go to the Ol Pejeta Conservancy, where the animals live, to collect egg cells from Fatu and Najin. These could then be fertilized with some frozen sperm and implanted into a surrogate mother, a southern white rhino (*Ceratotherium simum simum*).

But no one has ever made a viable rhino embryo using IVF, let alone implanted one into a surrogate, so the San Diego and Leibniz teams are each working to develop the technique in southern white rhinos, which number around 20,000. Hildebrandt is confident that obstacles such as implanting an embryo in a surrogate will be overcome in a matter of years. "Najin or Fatu will see another northern white rhino before they die. That I can guarantee," he says.

Najin and Fatu are currently the only source of egg cells for use in IVF. That limited gene pool means that it will not be possible to create a northern white rhino population that is sufficiently diverse to thrive in the wild. So in stage two, the researchers would try to reprogram frozen rhino cells into stem cells that have the capacity to develop into any type of tissue, including eggs and sperm (see 'Saving the northern white rhino'). In 2011, a team led by stem-cell scientist Jeanne Loring at the Scripps Research Institute in La Jolla, California, created such cells, known as induced pluripotent stem (iPS) cells from Fatu's skin cells (I. F. Ben-Nun *et al. Nature Meth.* **8,** 829–831; 2011). But generating sperm and eggs from iPS cells will not be simple, and could require rhino stem cells to be cultured alongside the reproductive tissue of other animals, such as mice. "All the technologies have been done but in other species," says Loring. "It's not certain these things are going to translate directly to rhinos."

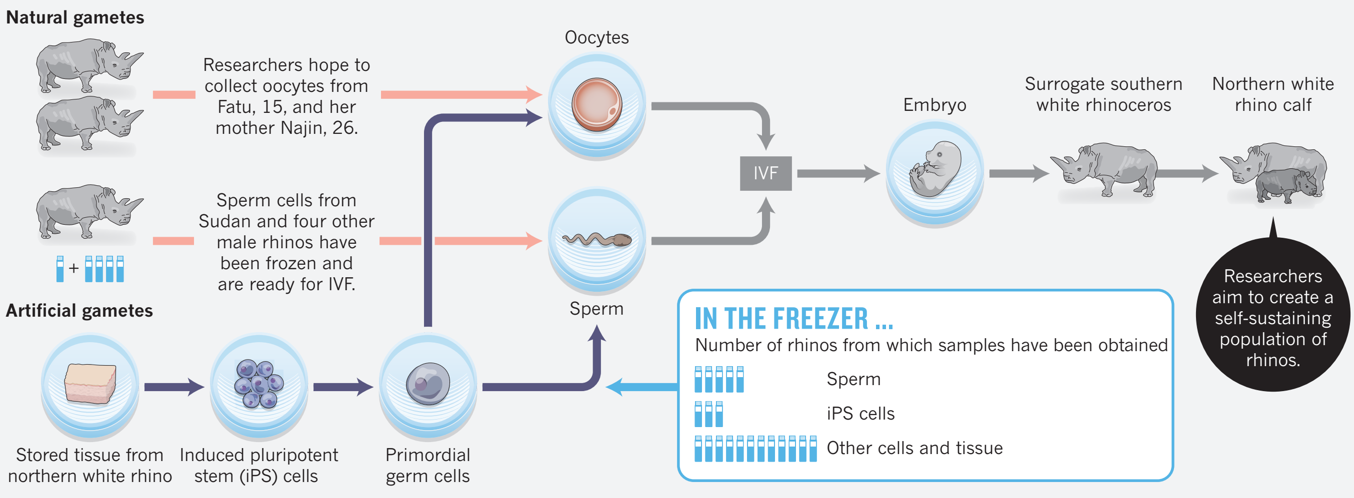
Conservationists have tried to bring species back from the brink using reproductive technologies before. In the 2000s, for instance, researchers attempted to use cloning to resurrect the Pyrenean ibex (*Capra pyrenaica*) and a species of wild ox (*Bos gaurus*). The cell reprogramming elements of the rhino plan are even more ambitious. "I don't see any technical deal-breakers," says George Church, a genome scientist at Harvard Medical School in Boston, Massachusetts. He hopes to use some of the same approaches to resurrect woolly mammoths, or at least engineer Asian elephants that can flourish in the Siberian steppe.

Funding could prove the greatest barrier. San Diego Zoo has raised around US$2 million for the project since its last northern white rhino, Nola, died last year; it declined to give an estimate of the project's total cost. Hildebrandt says that his team has had much less luck raising funds — and would need several million dollars to create a rhino through IVF.

Ryder says that the significant costs of rescuing and protecting northern white rhinos will be worth it — not only to save the species, but also to demonstrate what conservationists can do to rescue other animals. That precedent is

## SAVING THE NORTHERN WHITE RHINO

*Only three northern white rhinos are still alive, but Fatu, Najin and Sudan cannot breed naturally. So researchers plan to develop* in vitro *fertilization (IVF) and advanced cellular techniques to establish a viable population.*




what most worries Stuart Pimm, a conservation biologist at Duke University in Durham, North Carolina. "This says we can let species go to the very brink of extinction and modern technology can bring them back," he says. "There is a very substantial moral hazard in that."

"It's *Star Trek*-type science," says Michael Knight, chair of the International Union for Conservation of Nature's African Rhino Specialist Group. He worries that the effort could take away money from other rhino conservation efforts — including those directed at southern white rhinos, whose numbers are swelling thanks to good management, Knight says. "They should not be pushing this idea that they're saving a species. If you want to save a [rhino] species, put your money into southern white conservation." ■

# THE MATERIAL CODE

*Machine-learning techniques could revolutionize how materials science is done.*

BY NICOLA NOSENGO

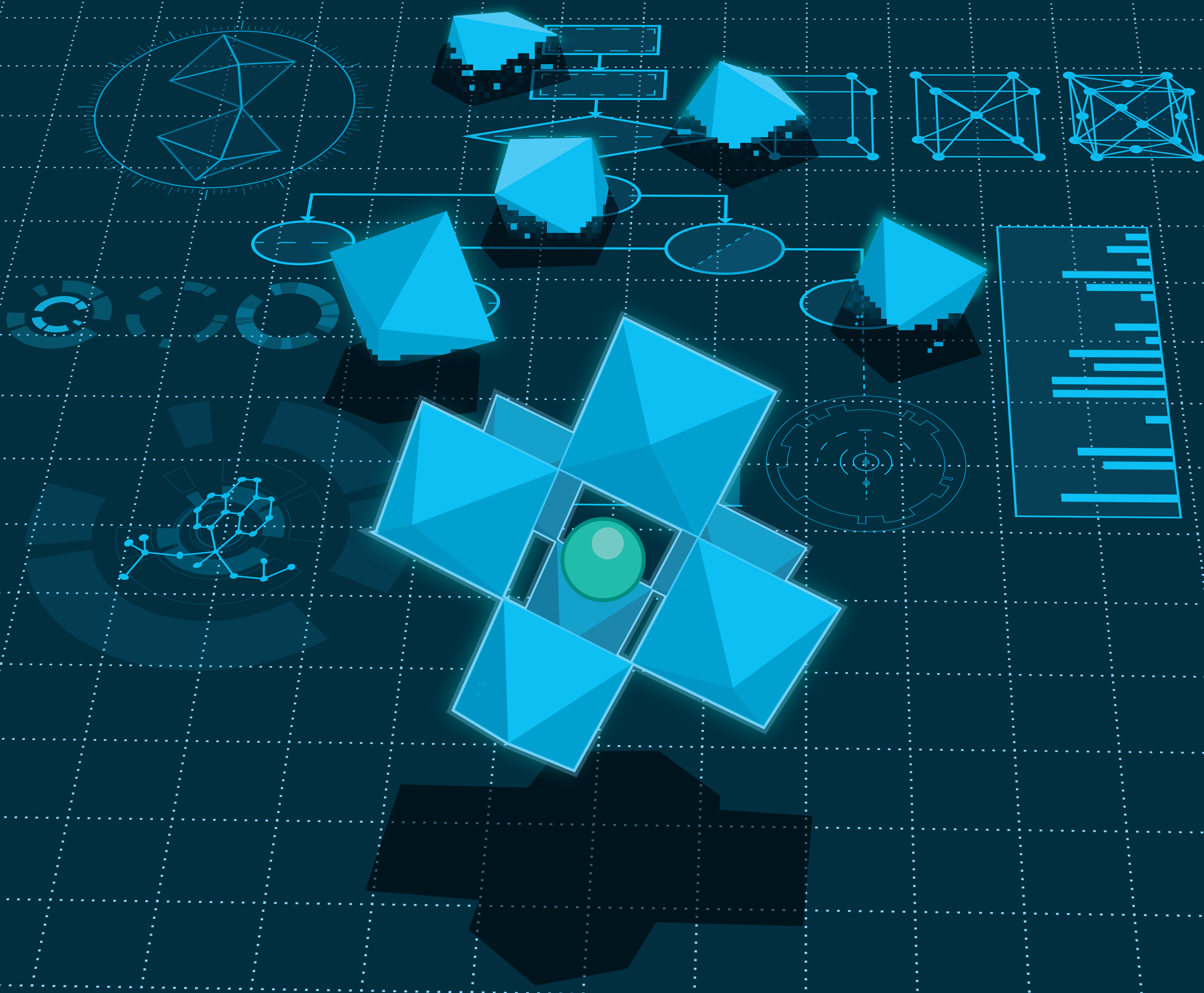


ILLUSTRATION BY NIK SPENCER/NATURE

It's a strong contender for the geekiest video ever made: a close-up of a smartphone with line upon line of numbers and symbols scrolling down the screen. But when visitors stop by Nicola Marzari's office, which overlooks Lake Geneva, he can hardly wait to show it off. "It's from 2010," he says, "and this is my cellphone calculating the electronic structure of silicon in real time!"

Even back then, explains Marzari, a physicist at the Swiss Federal Institute of Technology in Lausanne (EPFL), Switzerland, his now-ancient handset took just 40 seconds to carry out quantum-mechanical calculations that once took many hours on a supercomputer — a feat that not only shows how far such computational methods have come in the past decade or so, but also demonstrates their potential for transforming the way materials science is done in the future.

Instead of continuing to develop new materials the old-fashioned way — stumbling across them by luck, then painstakingly measuring their properties in the laboratory — Marzari and like-minded researchers are using computer modelling and machine-learning techniques to generate libraries of candidate materials by the tens of thousands. Even data from failed experiments can provide useful input[1]. Many of these candidates are completely hypothetical, but engineers are already beginning to shortlist those that are worth synthesizing and testing for specific applications by searching through their predicted properties — for example, how well they will work as a conductor or an insulator, whether they will act as a magnet, and how much heat and pressure they can withstand.

The hope is that this approach will provide a huge leap in the speed and efficiency of materials discovery, says Ceder, a materials scientist at the University of California, Berkeley, and a pioneer in this field. "We probably know about 1% of the properties of existing materials," he says, pointing to the example of lithium iron phosphate: a compound that was first synthesized[2] in the 1930s, but was not recognized[3] as a promising replacement material for current-generation lithium-ion batteries until 1996. "No one had bothered to measure its voltage before," says Ceder.

At least three major materials databases already exist around the world, each encompassing tens or hundreds of thousands of compounds. Marzari's Lausanne-based Materials Cloud project is scheduled to launch later this year. And the wider community is beginning to take notice. "We are now seeing a real convergence of what experimentalists want and what theorists can deliver," says Neil Alford, a materials scientist who serves as vice-dean for research at Imperial College London, but who has no affiliation with any of the database projects.

As even the proponents are quick to point out, however, the journey from computer predictions to real-world technologies is not an easy one. The existing databases are far from including all known materials, let alone all possible ones. The data-driven discovery works well for some materials, but not for others. And even after an interesting material is singled out on a computer, synthesizing it in a laboratory can still take years. "We often know better what we should be making than how to make it," says Ceder.

Still, researchers in this field are confident that there is a trove of compounds waiting to be discovered, which could kick-start innovations in electronics, energy, robotics, health care and transportation. "Our community is putting together a lot of different parts of the puzzle," says Giulia Galli, a computational materials scientist at the University of Chicago in Illinois. "And when they all click into place, materials prediction will become a reality."

"WE ARE NOW SEEING A REAL CONVERGENCE OF WHAT EXPERIMENTALISTS WANT AND WHAT THEORISTS CAN DELIVER."

## GENETIC INSPIRATION

The idea for this high-throughput, data-driven approach to materials discovery hit Ceder in the early 2000s, when he was at the Massachusetts Institute of Technology (MIT) in Cambridge and found himself inspired by the nearly completed Human Genome Project. "By itself, the human genome was not a recipe for new treatments," he says, "but it gave medicine amazing amounts of basic, quantitative information to start from." Could materials scientists learn some lessons from geneticists, he wondered. Could they identify a "materials genome" — Ceder's phrase — that encodes the properties of various compounds in the same way that biological information is encoded in DNA base pairs?

If so, he reasoned, that encoding must lie in the atoms and electrons that make up a given material, and in their crystal structure: the way they are arranged in space. In 2003, Ceder and his team first showed[4] how a database of quantum-mechanics calculations could help to predict the most likely crystal structure of a metal alloy — a key step for anyone in the business of inventing new materials.

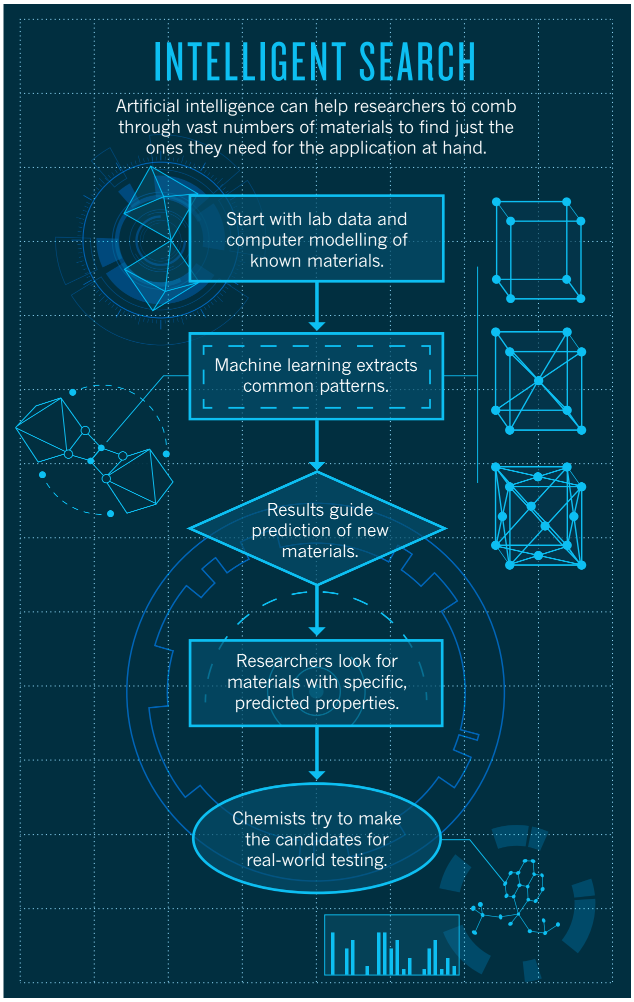
In the past, these calculations had been long and difficult, even for supercomputers. The machine had to go through an inordinate amount of trial and error to find the 'ground state': the crystal structure and electron configuration in which the energy was at a minimum and all the forces were in equilibrium. But in their 2003 paper[4], Ceder's team described a shortcut. The researchers calculated the energies of common crystal structures for a small library of binary alloys — mixes of two different metals — and then designed a machine-learning algorithm that could extract patterns from the library and guess the most likely ground state for a new alloy. The algorithm worked well, slashing the computer time required for the calculations (see 'Intelligent search').

"That paper introduced the idea of a public library of materials properties, and of using data mining to fill the missing parts," says Stefano Curtarolo, who that same year left Ceder's group to start his own laboratory at Duke University in Durham, North Carolina. The idea then gave birth to two separate projects. In 2006, Ceder started the Materials Genome Project at MIT, using improved versions of the algorithm to predict lithium-based materials for electric-car batteries. By 2010, the project had grown to include around 20,000 predicted compounds. "We started from existing materials and modified their crystal structure — changing one element here or another one there and calculating what happens," says Kristin Persson, a former member of Ceder's team who continued to collaborate on the project after she moved to the Lawrence Berkeley National Laboratory in California in 2008.

At Duke, meanwhile, Curtarolo set up the Center for Materials Genomics, which focused on research on metal alloys. Teaming up with researchers from Brigham Young University in Provo, Utah, and Israel's Negev Nuclear Research Center, he gradually expanded the 2003 algorithm and library into AFLOW, a system that can perform

calculations on known crystal structures and predict new ones automatically[5].

Researchers from outside the original group were getting interested in high-throughput computations as well. One such researcher was chemical engineer Jens Nørskov, who started using them to study catalysts for breaking down water into hydrogen and oxygen[6] while he was at the Technical University of Denmark in Lyngby, and later expanded the work as director of the SUNCAT Center for the computational study of catalysis at Stanford University in California. Another was Marzari, who was part of a large team developing Quantum Espresso: a program for quantum-mechanics calculations that was launched[7] in 2009. That is the code running on his mobile phone in the video.




NIK SPENCER/NATURE

#### MATERIALS GENOMICS

Still, computational materials science did not become mainstream until June 2011, when the White House announced the multimillion-dollar Materials Genome Initiative (MGI). "When people at the White House became familiar with Ceder's work they got very excited," says James Warren, a materials scientist at the US National Institute of Standards and Technology and executive secretary of the MGI. "There was a general awareness that computer simulations had got to the point where they could have a real impact on innovation and manufacturing," he says — not to mention the 'genomics' name, "which was evocative of something grand."

Since 2011, the initiative has invested more than US$250 million into software tools, standardized methods to collect and report experimental data, centres for computational materials science at major universities and partnerships between universities and the business sector for research on specific applications. But it is unclear how far this largesse has actually advanced the science. "The initiative brought a lot of good things, but also some re-branding," says Ceder. "Some groups started calling their research genomics this and genomics that, even though it had little to do with it."

One thing the MGI definitely did do, however, was to help Ceder and others realize their vision of an online database of materials properties. In late 2011, Ceder and Persson relaunched their Materials Genome Project as the Materials Project — having been asked by the White House to give up the 'genome' label to avoid confusion with the national effort. The following year, Curtarolo posted his own database, called AFLOWlib, based on the software he had developed at Duke[8]. And in 2013, Chris Wolverton, a materials researcher at Northwestern University in Evanston, Illinois, launched the Open Quantum Materials Database (OQMD)[9]. "We borrowed the general idea from the Materials Project and AFLOWlib," says Wolverton, "but our software and data are homegrown."

All three of these databases share a core of around 50,000 known materials taken from a widely used experimental library, the Inorganic Crystal Structure Database. These are solids that have been created at least once in a laboratory and described in a paper, but whose electronic or magnetic properties may have never been fully tested; they are the starting point from which new materials can be derived.

Where the three databases differ is in the hypothetical materials they include. The Materials Project has relatively few, starting with some 15,000 computed structures derived from Ceder's and Persson's research on lithium batteries. "We only include them in the database if we're confident the calculations are accurate, and if there is a reasonable chance that they can be made," says Persson. Another 130,000 or so entries are structures predicted by the Nanoporous Materials Genome Center at the University of Minnesota in Minneapolis. The latter focuses on zeolites and metal–organic frameworks: sponge-like materials with regularly repeating holes in their crystal structures that can trap gas molecules and could be used to store methane or carbon dioxide.

AFLOWlib is the largest database, featuring more than a million different materials and about 100 million calculated properties. That's because it also includes hundreds of thousands of hypothetical materials, many of which would exist for only a fraction of a second in the real world, says Curtarolo. "But it pays off when you want to predict how a material can actually be manufactured," he says. For example, he is using data from AFLOWlib to study why some alloys can form metallic glass — a peculiar form of metal with a disordered microscopic structure that gives it special electric and magnetic properties. It turns out that the difference between good glass formers and bad ones depends on the number and energies of unstable crystal structures that 'compete' with the ground state while the alloy cools down[10].

Wolverton's OQMD includes around 400,000 hypothetical materials, calculated by taking a list of crystal structures commonly observed in nature and 'decorating' them with elements chosen from almost every part of the periodic table[9]. It has a particularly wide coverage of perovskites — crystals that often display attractive properties such as superconductivity and that are being developed for use in solar cells as microelectronics. As the name suggests, this project is the most open of the three: users can download the entire database, not just individual search results, onto their computer.

All of these databases are works in progress, and their curators still spend a good share of their time adding more compounds and refining the calculations — which, they admit, are far from perfect. The codes tend to be quite good at predicting whether a crystal is stable or not, but less good at predicting how it absorbs light or conducts electricity — to the point of sometimes making a semiconductor look like a

metal. Marzari notes that even for battery materials, an area in which computational materials science is having its best success stories, standard calculations still have an average error of half a volt, which makes a lot of difference in terms of performance. "The truth is, some errors come with the theory itself: we may never be able to correct them," says Curtarolo.

Each group is developing its own techniques to adjust the calculations and make up for these systematic errors. But in the meantime they are already doing science with the data — and so are users from other groups. The Materials Project has identified several promising cathodes that may work better than existing ones in lithium batteries[11], as well as metal oxides that could improve the efficiency with which solar cells capture sunlight and turn it into energy[12]. And earlier this year, researchers from Trinity College Dublin used the AFLOWlib database to predict 20 Heusler alloys, a class of magnets that can be used for sensors or computer memories, and managed to synthesize two of them, confirming that their magnetic properties are very close to the predictions (see go.nature.com/v7djio).

"THE TRUTH IS, SOME ERRORS COME WITH THE THEORY ITSELF: WE MAY NEVER BE ABLE TO CORRECT THEM."

### EUROPEAN EXPANSION

Materials genomics has also crossed over to Europe — although usually by other names. Switzerland, for example, has created MARVEL, a network of institutes for computational materials science with the EPFL as its lead and Marzari as director. Using a new computational platform[13], he is creating a database called Materials Cloud that he is using to search for 'two-dimensional' materials, such as graphene, that are made from just a single layer of atoms or molecules. Such materials could be used in applications ranging from nanoscale electronics to biomedical devices. To find good candidates, Marzari is subjecting more than 150,000 known materials to what he calls 'computational peeling': calculating how much energy it would take to separate a single layer from the surface of an ordinary crystal. By the time the database is ready for public release later this year, he expects that preliminary runs will have yielded some 1,500 potential two-dimensional structures that can then be tested in experiments.

A few kilometres away in Sion, high in the Swiss Alps, computational chemist Berend Smit has set up another EPFL centre that develops algorithms for predicting hundreds of thousands of nanoporous zeolites and metal–organic frameworks. Other algorithms — including one that scans for certain pore shapes using techniques derived from facial-recognition software — then seek out the best candidates for absorbing carbon dioxide from the flues of fossil-fuel power plants[14].

Smit's work also shows that materials genomics can bring bad news. Many researchers had hoped to use nanoporous materials to build car tanks that could store more methane in less space. But after screening more than 650,000 computed materials, Smit's group concluded that most of the best ones have already been made[15]. New ones could bring only minor improvements, and energy targets currently set by US agencies — which bet on major technological improvements in methane storage — may be unrealistic.

As intriguing as these examples are, there are still many hurdles to overcome before materials genomics can live up to its promises. One of the largest is that computer simulations still give few clues on how an interesting material can be made in a lab — let alone mass produced. "We come up with interesting ideas for new compounds all the time," says Ceder. "Sometimes it takes two weeks to make it. Other times we still can't make it after six months, and we don't know whether we haven't done the right thing, or it just can't be made."

Both Ceder and Curtarolo are trying to develop machine-learning algorithms to extract rules from known manufacturing processes to guide the synthesis of compounds.

Another limitation is that materials genomics has been hitherto applied almost exclusively to what engineers call functional materials — compounds that can perform a task such as absorbing light in a solar cell or letting electrical current pass in transistor. But the technique does not lend itself well to studying structural materials, such as steel, that are needed to build, for example, aircraft wings, bridges or engines. This is because mechanical properties such as a material's springiness and hardness depend on how it is processed — something that quantum-mechanical codes by themselves can not describe.

Even in the case of functional materials, current computer codes work well only for perfect crystal structures — which are only a small part of the materials realm. "The most interesting materials of the future will probably be assembled at the microscopic level in creative ways," says Galli. They may be assemblies of nanoparticles, crystals with strategically placed defects in their structures, or heterogenous materials made by intertwining different compounds and phases. To predict such materials, says Galli, "you need to calculate many properties at once and how the system will evolve in time and at specific temperatures". There are methods to do that, she says, "but they are still too computationally expensive to be used in high-throughput studies".

In the short term, more data exchange with experiments can give computations a reality check and help to refine them. To that end, Ceder is working with a group at MIT on software that reads papers in experimental materials science and automatically extracts information on crystal structures in a standard format. "We plan to begin adding these data to the Materials Project in a few months," he says.

And in the long run, some help will come from Moore's law: as computational power continues to increase, some techniques that are out still of reach for current computers may soon become viable.

"We've moved away from the artisanal era of computational materials science, and into the industrial phase," says Marzari. "We can now create assembly chains of simulations, put them to work, and explore problems in totally new ways." No computationally predicted material is on the market just yet. "But let's talk again in ten years," says Galli, "and I think there will be many." ■

**Nicola Nosengo** *is a freelance writer based in Rome.*


1. Raccuglia, P. *et al. Nature* **533,** 73–76 (2016).
2. Björling, C. O. & Westgren, A. *Geol. Fören. Stock. För.* **60,** 67–72 (1938).
3. Padhi, A. K., Nanjundaswamy, K. S. & Goodenough, J. B. *J. Electrochem. Soc.* **144,** 1188–1194 (1997).
4. Curtarolo, S., Morgan, D., Persson, K., Rodgers, J. & Ceder, G. *Phys. Rev. Lett.* **91,** 135503 (2003).
5. Curtarolo, S. *et al. Comput. Mater. Sci.* **58,** 218–226 (2012).
6. Greeley, J., Jaramillo, T. F., Bonde, J., Chorkendorff, I. & Nørskov, J. K. *Nature Mater.* **5,** 909–913 (2006).
7. Giannozzi, P. *et al. J. Phys. Condens. Matter* **21,** 395502 (2009).
8. Curtarolo, S. *et al. Comput. Mater. Sci.* **58,** 227–235 (2012).
9. Kirklin, S. *et al. npj Comput. Mater.* **1,** 15010 (2015).
10. Perim, E. *et al.* Preprint at http://arxiv.org/abs/1601.08233 (2016).
11. Jain, A., Shin, Y. & Persson, K. A. *Nature Rev. Mater.* **1,** 15004 (2016).
12. Castelli, I. E. *et al. Adv. Energy Mater.* **5,** 1400915 (2015).
13. Pizzi, G., Cepellotti, A., Sabatini, R., Marzari, N. & Kozinsky, B. *Comput. Mater. Sci.* **111,** 218-230 (2016).
14. Lin, L.-C. *et al. Nature Mater.* **11,** 633–641 (2012).
15. Simon, C. M. *et al. Energy Environ. Sci.* **8,** 1190–1199 (2015).

# Silent no more

ILLUSTRATION BY GARY NEILL

## After being muzzled for nine years, government scientists in Canada are now allowed to speak out about their work.

By ***Lesley Evans Ogden***

Early one Thursday morning last November, Kristi Miller-Saunders was surprised to receive a visit from her manager. Miller-Saunders, a molecular geneticist at the Canadian fisheries agency, had her reasons to worry about attention from above. On numerous occasions over the previous four years, government officials had forbidden her from talking to the press or the public about her work on the genetics of salmon — part of a broad policy that muzzled government scientists in Canada for many years. At one point, a brawny 'minder' had actually accompanied her to a public hearing to make sure that she didn't break the rules.

But the meeting last autumn was different. Miller-Saunders' manager at Fisheries and Oceans Canada (DFO) in Nanaimo walked in with a smile and gave her advance notice that the newly elected government would be opening up scientific communication: she and other federal researchers would finally be free to speak to the press. "It was like a weight was being lifted," she says. Important findings on climate change, depletion of the ozone layer, toxicology and wildlife conservation that had been restricted for so long could now be openly discussed.

Canadian scientists celebrated the move far and wide. Shark researcher Steve Campana danced in his office at the University of Iceland in Reykjavik, where he had relocated after leaving the DFO because of the communications constraints and other limitations.

Six months later, the government is loosening its grip on communications but the shift at some agencies has not been as swift and comprehensive as many had hoped. And with the newfound freedom to speak, the full impact of the former restrictions is finally becoming clear. Canadian scientists and government representatives are opening up about what it was like to work under the former policy and the kind of consequences it had. Some of the officials who imposed the rules are talking about how the restrictions affected the morale and careers of researchers. Their stories hint at how governments control communications in even more politically repressive countries such as China, and suggest what might happen in Canada if the political winds reverse.

"It was not a good time for journalists. It was not a good time for scientists. It was not

a good time for morale in the federal community, and it was not a good time for Canadian citizens," says Paul Dufour, a science-policy analyst at the University of Ottawa.

## Set to silence

The crackdown on government scientists in Canada began in 2006, after Stephen Harper of the Conservative Party was elected prime minister. During the nine-year Harper administration, the government placed a priority on boosting the economy, in part by stimulating development and increasing the extraction of resources, such as petroleum from the oil sands in Alberta. To speed projects along, the administration eased environmental regulations. And when journalists sought out government scientists to ask about the impacts of such changes, or anything to do with environmental or climate science, they ran into roadblocks.

For decades before the Harper administration, reporters had been free to call up government researchers directly for interviews. But suddenly, all requests for interviews had to be sent to government communications offices, which then had to get approval from multiple tiers of bureaucrats higher up. "It was an incredible rigmarole to try and get the most innocuous bit of information to media or the public," says Diane Lake, who was a communications officer with the DFO at the time.

Lake had been a newspaper reporter for a dozen years before joining the department in 1992, so she knew what journalists needed to produce stories. She has fond memories of her time as a communications officer before the Harper years, but after he took office, her job became less about communicating science and more about censoring it. When journalists called her trying to reach scientists, she was required to get approval for scripted answers that researchers could give, but she found the authorization process opaque and arbitrary. "There were never any written protocols on what would pass muster and what wouldn't," she says. "I would always say, 'can you write that down?' to folks in Ottawa." No one ever did.

Because the scripts had to be endorsed by "legions of approvers" in a convoluted process, meeting reporters' deadlines was "kind of hopeless", says Lake. The starkest example for her came in 2011, when Miller-Saunders (then Miller) and her colleagues published a paper in *Science* that investigated why unusual numbers of sockeye salmon (*Onchorhynchus nerka*) were dying in British Columbia's Fraser River on their way to spawn (K. M. Miller *et al. Science* **331,** 214–217; 2011). Through genomic analysis, the researchers found evidence that a virus might be to blame. The topic was sensitive in part because some scientists and environmentalists had previously raised concerns that fish farms could transfer diseases to wild salmon.

*Science* had alerted journalists about the paper days ahead of its publication under an embargo, giving reporters time to conduct interviews and write their stories. Many journalists had contacted Lake with requests to speak with Miller-Saunders, and Lake had been busy setting up interviews during the days before publication. But the permission process dragged on, and Lake and Miller-Saunders had to postpone those interviews repeatedly.

Then, on the day of the paper's publication — 14 January — Lake got word from Ottawa that Miller-Saunders had been denied permission to talk to reporters at all. "Obviously, journalists were very upset, and it sort of snowballed from there," Lake says. Many reporters wrote stories about the muzzling of a government scientist rather than about the genetics of salmon.

> **"It was like an *iron curtain* was drawn across communicating research to *Canadians*."**

Journalists who wanted interviews with Miller-Saunders were told to contact her co-authors outside the government. "The unfortunate thing was that my co-authors were not genomic scientists," Miller-Saunders says, so they couldn't readily address specific questions about the genetic aspects of the study.

The "Kristi Miller debacle", as Lake calls it, was just one high-profile example of scientists being silenced. But there were hundreds of others, she says. "It was like an iron curtain was drawn across communicating research to Canadians."

The federal government maintained that it was inappropriate for Miller-Saunders to speak to reporters because she was part of a judicial enquiry into the management of sockeye salmon, known as the Cohen Commission. At a public enquiry of the commission in 2011, the DFO assigned Miller-Saunders a media officer and a bodyguard, whom Miller-Saunders describes as a "very nice burly man". Miller-Saunders was kept in a separate room, away from the media and public, when not testifying. Her husband and daughter were there with her. "It was all very friendly and meant to keep me from distraction and being a distraction," she says. Because she was not permitted to speak for herself, a media officer answered all questions on behalf of Miller-Saunders. "It was all a very surreal experience," she says. University scientists on the commission, by contrast, could freely speak to the media freely.

The decision to muzzle Miller-Saunders was clearly political, says Calvin Sandborn, legal director of the University of Victoria's Environmental Law Centre. "There are all sorts of enquiries where experts talk about their findings outside of the hearing room."

Although the approval 'rules' were unwritten, Lake says it became clear over time what stories were likely to be permitted. Under Harper, government-science stories, "could only reflect economics, and what you could sell, not what you could save or conserve", she says.

Lake's work environment became a culture of frustration, low morale and fear, she says. Midway through the Harper years, she attended a meeting called by the DFO's Pacific-region director-general, Paul Sprout. Lake says that Sprout was "fair, and treated staff with integrity". But on this occasion, "he told staff they were not to speak critically about the Harper government, even on their own time".

That atmosphere eventually wore Lake down. She retired several years early, in 2013, explaining that she found the atmosphere at work "untenable". Now, she spends her time writing, volunteering and working in a community garden. She would like to have served in Canada's new government, she says, in a communications role "where public employees can actually do their job".

Sprout, now retired from his 34-year career with the DFO, denies having said that employees had to wait until they left their posts before saying anything critical about the government. He confirms, however, that the DFO's policy was "unequivocal that any approval for doing media interviews would have to be approved by the director-general of communications", who was based in Ottawa.

Sprout says that it was his responsibility to enforce the policy so that communications employees and scientists in his department would not face any repercussions in their personal careers. "I had to make sure that the policies of the department were respected. That was my job," he says.

When he started out as a fisheries biologist in the late 1970s, there was much more flexibility in communications, even when other Conservative governments were in power, Sprout says. During the Harper era, "there were a lot of limitations on being able to speak", says Sprout. "It was difficult to actually get media interviews, even when we wanted to encourage them."

## Toxic environment

Not all scientists were willing to comply with Canada's closely controlled communications practices. One senior scientist who flouted the rules was Robie Macdonald, a biogeochemical oceanographer who was at the DFO's Institute of Ocean Sciences (IOS) in Sidney. He started his career with the DFO in 1973, and had worked under many federal governments.

Early in his career, there was no written


JOHN LEHMANN/GLOBE AND MAIL/CANADIAN PRESS

**Kristi Miller-Saunders was not allowed to talk to the press about her work on salmon management.**

media policy, but scientists understood that "they should comment on science and science issues and shouldn't comment on policy", he says. The Harper government, however, "made the process so cumbersome that most media people would not bother talking to you to start with".

Macdonald's group studied ocean contaminants, and the researchers ran afoul of the administration because they often identified environmental problems, such as the toxic effects of mercury and persistent organic pollutants on wildlife. Under Harper, contaminants research was removed from the DFO's mandate and toxicologists were fired or transferred, he says. When Macdonald's work on contaminants was cancelled, he retired early to continue his research, unpaid.

Another federal scientist who retired earlier than he had intended — in part because of media muzzling — was Ian Stirling, a prominent biologist with Environment and Climate Change Canada, the federal department that conducts research in areas including air quality, ozone, climate, weather, pollution and wildlife. Stirling began studying polar bears in 1970, but such research attracted scrutiny under the Harper government because scientists had shown that the animals were sensitive to climate change and the loss of sea ice.

Stirling says that the policies during the Harper administration reminded him of a another regime that had tight control over the media. During the 1970s, he had gone to meetings in Canada that were also attended by Soviet scientists. The visiting researchers would arrive, he says, "with a KGB guy, who would stand there with no smiles, a scowl on his face and arms crossed". Stirling still finds it unbelievable that the Canadian government used similar tactics at conferences. In 2012, for example, the Canadian news outlet CBC reported that media minders had shadowed scientists from Environment Canada at a meeting of the International Polar Year in Montreal.

Some officials say that the situation was not as bad as it has been portrayed. One manager within Environment Canada spoke to *Nature* on condition of anonymity. He says that the "muzzling" label used by the media is an over-exaggeration. "I think that's a bit of a coarse way to articulate it. What was done really was a bit more nuanced than that," he says. The vetting process required approval from such a high level "that the probability of getting that within a very tight, and very common, media timeline, wasn't great", he says.

**"It was not a good time for *journalists.* It was not a good time for *scientists.*"**

"Sometimes we got approval, and sometimes we didn't. It wasn't always clear why," he says. Sometimes even stories about good news wouldn't get approved. He attributes this to the sheer volume sent "into the black box of decision-making". The most profound effect, he says, was that "people on both sides stopped trying".

Now, the manager says, media protocols in his office are "back to more or less the old way of doing it". If a journalist contacts one of his scientists directly, the researcher can do an interview but is required to inform a manager and communications officer beforehand. That's progress, but it offers less freedom than the DFO's new directive that scientists can now talk to media first, and let communications staff know later.

## Government crackdown

Some departments are clearly struggling with the transition, as *Nature* found when it requested current media protocols for scientists from several government departments. Parks Canada provided information that had been published in 2006 and was updated in 2012, during the Harper administration. Canadian journalists continue to report difficulties in setting up media interviews with Parks Canada scientists.

Some scientists and communications staff worry that a shift in the political winds could bring back restrictive policies. "It's hard to say that it wouldn't happen again. It happens all over the world in totalitarian governments," Lake says.

A former journalist from China says that scientists there are censored, but that the restrictions are often lighter than those imposed on other sectors because science is considered ideologically free and the state censorship agency may not have the capacity to censor every researcher. But he also says that scientists there are generally reluctant to give interviews. "Scientists in China are not accustomed to talking to journalists," he says.

The muzzling of scientists is an ongoing concern even in some of the most open countries. The Union of Concerned Scientists (UCS) in Cambridge, Massachusetts, started tracking the issue in the United States during the administration of President George W. Bush, when government scientists complained that their data were being altered or suppressed and that they were unable to talk to the media. When President Barack Obama took office in 2009, he vowed to end such practices and ordered government departments to adopt scientific-integrity policies; but journalists and scientists still report problems with some agencies.

Gretchen Goldman, the lead analyst with the UCS on this issue, says that one thing Canada might learn from the US experience is that it takes time for a culture of transparency to take root. Even after a more open administration assumes power, many staff members remain from the previous government, and have been trained in the more-restrictive policies. "Practices often lag the policy," she says.

It could take years for Canadian scientists to recover from heavy funding cuts, low morale and tight control over communication. Looking back over what happened, Macdonald remembers something his grandmother once told him. "It takes ten years to make a good garden, but you can wreck it in six months," he says. "It's like that with science." ■

**Lesley Evans Ogden** *is a journalist in Vancouver, Canada.*

Specialists in Peru fumigate a cemetery in an effort to prevent Chikungunya and Zika viruses from spreading.

ERNESTO BENAVIDES/AFP/GETTY

# Security spending must cover disease outbreaks

**Tadataka Yamada**, **V. Ayano Ogawa** and **Maria Freire** call for research and development funding and coordination to counter global infectious-disease threats.

The health emergency precipitated by the Zika virus is a salutary reminder: global preparedness for emerging pathogens with endemic or pandemic potential is crucial and needs an overhaul. These crises are not rare — Lassa fever, Ebola virus, Middle East respiratory syndrome, H1N1 influenza and severe acute respiratory syndrome (SARS) have surfaced in head-spinning succession over the past 10–15 years. Each emergence proves how woefully unprepared the global community is to deal with worldwide health emergencies that have deep societal and economic impact.

Diagnostic tools, medicines and vaccines are in limited supply, non-existent or too costly — many people die and many more suffer in each outbreak as a result. Fear and panic spread, borders are closed, travel is restricted and commerce is shut down. After the recent Ebola outbreak in West Africa, the direct financial repercussions on Liberia, Sierra Leone and Guinea could amount to around 10%[1] of the nations' gross domestic product for 2014–15; the cost of SARS to the global economy in 2003, exceeded US$40 billion[2].

The health, economic and social consequences of a global health emergency are as great a threat to global and national security as those of terrorist actions. Although the world has gone to great expense and effort to prepare for the latter, it has done unacceptably little to prepare for the former, given the solemn responsibility of nations to ensure the health and security of their citizens. The United States spends at least $100 billion a year on counterterrorism efforts; it invests just $1 billion on pandemic and emerging infectious-disease programmes[3].

In this context, the Commission on ▶

a Global Health Risk Framework for the Future — an independent, international panel — published recommendations in January for addressing future global infectious-disease threats[4]. The 17-member commission has a secretariat at the US National Academy of Medicine. It was supported by seven private donors as well as the US Agency for International Development, and sought advice from more than 200 global technical experts from government, private industry, academia, non-governmental organizations and foundations. The commission's report addressed pandemic preparedness from four perspectives: governance, health systems, financing and research and development (R&D).

Here we expand on the R&D element of these recommendations. Several excellent global proposals and initiatives have arisen in the past year that are relevant to R&D for pandemic preparedness. One is a proposal to create a fund to support vaccine development. Another is an R&D Blueprint, issued by the World Health Organization (WHO), which aims to implement a road map for R&D preparedness for known priority pathogens and to facilitate roll out of an emergency R&D response in a timely manner for emerging ones. But gaps remain — conceptually, practically and financially — and these need to be plugged, urgently, in the following ways.

## MORE FUNDS

Society — national governments, industry, charities and others — needs to invest an extra \$1 billion per year for 15 years, over and above the amount currently being spent on R&D for infectious diseases and global preparedness. This is equivalent to the R&D budget of a medium-sized pharmaceutical company with a portfolio of products in various stage of development (see go.nature.com/4hfdrj).

These funds would be used in three ways: in the targeted expansion or acceleration of ongoing R&D projects (excluding those that address antimicrobial resistance, which deserve their own targeted funds and efforts); for the development of core functions, such as clinical-trial infrastructure and manufacturing capacity; and to spur innovation, especially in new platforms that could allow 'plug and play' strategies, offering the potential to move quickly from the identification of a pathogen to the development and manufacturing of a product.

Is \$1 billion too much or too little? It is less than 2% of the United States' annual budget for homeland security[5] and less than 0.2% of its defence budget[6]. Thus it is in our view a reasonable, attainable sum. Some feel that there is little enthusiasm from funders, including governments, for extra pooled resources for R&D — but jump-starting the enterprise is paramount. The goal is the security of the world's population.

We contend that the money should come from multiple sources, including national-security and defence budgets. In non-emergency times, governments must support the training of scientific and medical personnel to carry out basic-research activities and provide them with adequate local laboratories in which to work. Poorer countries have smaller budgets, but health should be their top priority. This strategy is akin to basic military preparedness, requiring resources, practice, vigilance and long-term commitment.

Other crucial contributors to preparedness include private industry, particularly pharmaceutical and biotechnology companies, foundations, charities and, importantly, non-traditional actors such as insurance companies and other funders. WHO director-general Margaret Chan has noted that the pharmaceutical industry spent almost \$1 billion to develop Ebola vaccines in the past two years without any return on investment[7].

To attract and retain more private-sector involvement in R&D, national governments and foundations must put in place reasonable incentives. This is key for conditions with uncertain markets or low financial returns. One such lever is the priority-review voucher that may be issued in the United States by the Food and Drug Administration to those who develop treatments for diseases that typically do not command big commercial markets, such as river blindness (onchocerciasis).

*"We must not repeat the events of the H1N1 influenza pandemic in 2009."*

In recent years, philanthropic foundations have played an increasingly important part in R&D for global health. Organizations such as the Bill & Melinda Gates Foundation and Médecins Sans Frontières (also known as Doctors Without Borders) have funded new mechanisms for drug and vaccine development for conditions including tuberculosis, malaria, dengue fever, leishmaniasis and Chagas disease. Such operations, known as product development partnerships (PDPs), decouple basic-research expenditures and the cost of failure. PDPs are a powerful mechanism to address product gaps, provided that the basic biology of disease is understood and a path for development is identified.

## MORE COORDINATION

To build and expand on independent public and private-sector activities and ensure synergy, we propose the creation of an independent high-level expert committee. It would help to coordinate research activities, prioritize investments, monitor progress, minimize duplication of effort and make timely decisions. This 15-member Pandemic Product Development Committee (PPDC) would help to make the best of scarce new resources, such as the ability to carry out clinical trials on the ground. It would not undertake direct management of any specific project or have decision-making authority over activities and budgets in ongoing research efforts.

The chair of this committee would be appointed by the WHO director-general following broad consultation with the key stakeholders. The chair and the members of the committee, who would be supported by a small, expert secretariat at the WHO, must have extensive knowledge and experience in the discovery, development, regulatory review and manufacture of medical products and related technologies. The PPDC should feature representatives from industry, academia, the civil service and society. The chair would be a standing member of and accountable to an independent technical governing board, proposed by the commission to oversee the global pandemic preparedness effort[4].

This governance model has ties to, but is separate from, the WHO. The proposal is based on several factors, including the WHO's global responsibility for health emergencies, the need to tap multiple R&D parties and the importance of providing the highest level of technical expertise in a neutral forum. Making the PPDC fully part of a United Nations agency would limit the flexibility required for rapid decision-making. Divorcing the PPDC completely from the WHO would undermine the agency's leadership role in health emergencies.

## MORE ENGAGEMENT

During a crisis there is an understandable urge to try unproven technologies on people who are certain to die unless something is done. Yet it is only by maintaining a commitment to scientific rigour that the world has medicines that cure, and vaccines that prevent, disease. Efforts to create new treatments, including those for infectious diseases, must include randomized clinical trials despite the challenges, unless there is some other scientifically valid approach that could lead to similarly actionable information.

Under these circumstances, it is evident that the communities in which trials are being conducted and where resultant products will be distributed must be involved in any R&D effort from the start. Only by understanding their role as partners in the research effort and the societal benefit of their participation in a placebo-controlled trial will clinical-trial volunteers be able to understand and accept

JOHN MOORE/GETTY

Vaccine trials were conducted during West Africa's devastating Ebola outbreak.

## PANDEMIC PREPAREDNESS

### *Six steps*

- **Negotiate trial designs**, including protocols for clinically testing different products against one control group.
- **Agree sharing policies** for reagents, data, patents and other intellectual property.
- **Agree regulatory policy**, including standards for reviewing and approving products for emergencies, and roles between and within drug agencies in affected countries.
- **Design liability protection** for those who conduct the research and development and for compensation to people affected by unexpected events resulting from experimental interventions.
- **Prioritize allocation of resources** such as candidate compounds, instrumentation or clinical-trial sites.
- **Ensure capacity** for rapid manufacturing, strategic stockpiling and prompt delivery of products.

the risks. The communication of crucial information regarding clinical studies will often require the engagement of trusted community or religious leaders and translation into native languages to establish understanding and trust.

Before a crisis, it is the responsibility of all those involved in infectious-disease research and development — public and private — to ensure that drug and vaccine candidates can quickly move forwards. Preparedness must encompass six key activities (see 'Six steps') which should be discussed by the PPDC and implemented by the appropriate stakeholders.

Crucially, stakeholders must agree that the fruits of these efforts will be distributed first to those in greatest need or at greatest risk. We must not repeat the events of the H1N1 influenza pandemic in 2009. Nations with manufacturing facilities distributed vaccines domestically before exporting them; some wealthy nations without vaccine-manufacturing capacity paid substantial sums to reserve the remaining supply. Meanwhile, robust modelling studies indicated that more than 90% of the deaths from a potential influenza pandemic would probably occur in the world's poorest countries[8].

## ACT NOW

R&D for products to address emerging health threats is severely limited and fragmented. Substantial investment and a global commitment are needed to better coordinate independent activities. Components of the basic arsenal such as fit-for-purpose medicines, vaccines, diagnostics and personal protective equipment must exist so that first responders and medical personnel can identify, treat and contain an outbreak.

At the global level, countries must ensure a coordinated, nimble R&D response to health outbreaks. This should include: the comprehensive search for and assessment of existing technologies to tackle the disease; the testing of candidate drugs and vaccines that can be put quickly into development; the repurposing of existing technologies; and worldwide manufacturing capacity that is ready for the rapid production of high-quality drugs and vaccines.

To be clear, the funds we call for are to increase the current worldwide R&D expenditures, not to replace them. Naturally, basic research into the aetiology of disease and the biology that underpins diseases with pandemic potential must be strongly supported by governments, industry and foundations. Such work is the foundation on which new life-saving tools will be built.

Three principles should guide R&D for epidemic or pandemic disaster preparedness. First, we must maintain consistently high ethical and scientific standards, particularly during crises. Second, we must define protocols and approaches to engage local scientists and community members early in the conduct of research. And third, we must agree on ways to expedite medical-product approval, manufacture and distribution.

It is imperative that these recommendations are adopted on a global scale. There will be many reasons why some may argue with one or more, and there may be a temptation to delay or forgo the necessary commitments. But we must act. We cannot afford to lose this battle. ■

**Tadataka Yamada** *is a commissioner on the Global Health Risk Framework Commission, and venture partner at Frazier Healthcare Partners, Seattle, Washington, USA.* **V. Ayano Ogawa** *is associate programme officer for the Global Health Risk Framework Commission at the US National Academy of Medicine, Washington DC, USA.* **Maria Freire** *is a commissioner on the Global Health Risk Framework Commission, and president and executive director of the Foundation for the National Institutes of Health, Bethesda, Maryland, USA.*
*e-mail: tachi.yamada@gmail.com*

1. United Nations Development Group, Western and Central Africa. *Socio-Economic Impact of Ebola Virus Disease in West African Countries: A Call for National and Regional Containment, Recovery and Prevention* (United Nations Development Group, 2015).
2. Lee, J. W. & McKibbin, W. J. in *Learning from SARS: Preparing for the Next Disease Outbreak: Workshop Summary* (eds Knobler, S. *et al.*) 92–109 (Natl Acad. Press, 2008).
3. Boddie, C., Sell, T. K. & Watson, M. *Health Secur.* **13,** 186–206 (2015).
4. Commission on a Global Health Risk Framework for the Future. *The Neglected Dimension of Global Security: A Framework to Counter Infectious Disease Crises* (Natl Acad. Press, 2016).
5. US Department of Homeland Security. *Budget-in-Brief* (2015).
6. Office of the Under Secretary of Defense (Comptroller) Chief Financial Officer. *United States Department of Defense Fiscal Year 2016 Budget Request: Overview* (2015); available at http://go.nature.com/vxnlp4
7. Ward, A. 'WHO chief wants help for drug industry to fight global pandemics' *Financial Times* (2016).
8. Murray, C. J. L., Lopez, A. D., Chin, B., Feehan, D. & Hill, K. H. *Lancet* **368,** 2211–2218 (2006).

T.Y. declares competing financial interests: see go.nature.com/chvdjg for details.

PHOTO RESEARCHERS/FLPA

Bees, including *Apis mellifera* (pictured), perform a waggle dance to tell others about new resources.

ETHOLOGY

# Intrepid translator of the hive

**Mark L. Winston** reviews a study of Karl von Frisch, the ethologist who unravelled bee communication.

One of the most remarkable scientific discoveries of any century was honeybee dance language. Foragers and scouts run and turn to communicate the distance, direction and quality of flowers or nest sites to other worker bees. Many scientists were involved in elucidating the dance's sophisticated communicative functions, but Austrian ethologist Karl von Frisch (1886–1982) delivered the main results during the 1940s, for which he won the 1973 Nobel Prize in Physiology or Medicine. Excellent observations, painstaking experimental designs, laborious research and some controversy made von Frisch's work novelistic in its drama. Brilliance was required to discover and translate the language of an invertebrate as behaviourally complex as the bee.

The story has been well told in a number of books, most notably von Frisch's own 1967 classic, *The Dance Language and Orientation of Bees* (Harvard University Press). Now, in *The Dancing Bees*, Tania Munz gives us von Frisch the man, whose stellar accomplishments are well known but whose personal history has not been so well described — especially his years under the Third Reich.

Many German scientists fled the country when Hitler came to power; those who remained were expected to contribute their expertise to the war effort. Although von Frisch was never a member of the Nazi Party, his research flourished against the odds during the Second World War, while he was based at the Zoological Institute at the University of Munich.

As Munz rivetingly shows, von Frisch was triply vulnerable. His maternal grandmother was deemed Jewish under Nazi doctrine. His laboratory reputedly employed numerous Jewish researchers, although Munz does not address the accuracy of those claims. And von Frisch had enemies in academia, driven by either professional jealousy or rabid anti-Semitism. They included astronomer Wilhelm Führer, head of the University of Munich's Instructor's League, and botanist Ernst Bergdolt, president of the National Socialist Lecturer's League — both members of the Nazi Party. Yet von Frisch also had supporters. Among them were two powerful names in German science: Alfred Kühn and Fritz von Wettstein, both from Berlin's Kaiser Wilhelm Institute for Biology. They lobbied hard on his behalf. But in the end, it was the bees that earned him an academic reprieve. In 1941, *Nosema*, a dysentery-causing fungal parasite, destroyed 800,000 of Germany's bee colonies, threatening the regime's already strained agricultural productivity.

Von Frisch was tasked to address this problem. He interpreted this to include devising ways to attract bees to crops, a topic that led him towards the discovery of dance language. He had described the dances as early as 1927 (in a book also entitled *The Dancing Bees*; Springer) without understanding their remarkable functions. But his wartime research revealed their importance. Most significantly, he found that forager bees communicated the distance and direction to flower sources through a 'waggle' dance. The bees make a straight run while waggling and buzzing: the duration indicates distance, and the angle of the dance on the comb relative to the vertical indicates direction relative to the Sun.

Von Frisch's ability to block out the chaos around him was astounding. His research output was prodigious, even as his institute and lab were reduced to rubble, food supplies dwindled and friends, colleagues and relatives were wounded or killed.


**The Dancing Bees: Karl von Frisch and the Discovery of the Honeybee Language**
TANIA MUNZ
*University of Chicago Press: 2016.*

Munz also covers von Frisch's postwar research, when he further demonstrated his extraordinary capacity to observe, design experiments and recognize the paradigm-breaking significance of data. His work on the colour vision of bees set experimental protocols that are still used by behavioural scientists, and towards

the end of his career, he uncovered the mystery of how bees in flight orient themselves to the Sun's position using polarized light.

Not all of *The Dancing Bees* is spellbinding. Munz devotes a chapter to dance-language denial, a controversy based on meagre evidence that, with the benefit of historical hindsight, deserves a couple of paragraphs at most. A few inserted vignettes about honeybee observations in the eighteenth century by Swiss naturalist François Huber, and von Frisch's films about fish behaviour, interrupt the book's flow. Moreover, key aspects of von Frisch's personal life are under-represented. His relationships with his wife, children and friends are mentioned, but further elaboration would have enriched our understanding of how he persevered through scientific controversy and historical tragedy.

Von Frisch clearly did not collaborate in any substantial way with the Nazis. Munz is largely silent on whether he could or should have been more proactive or outspoken against the regime. She writes: "It is difficult

JACK GAROFALO/PARIS MATCH ARCHIVE/GETTY

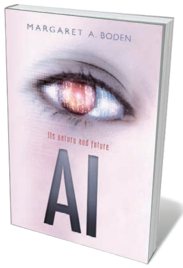
**Karl von Frisch translated the bees' dance.**

to shake the image of a scientist who escaped the horrors that surrounded him by burying himself in his work." After the war, von Frisch wrote that "many professors welcomed the changes, some out of caution, others from conviction. And soon it was clear that any serious opposition would lead to one's personal destruction."

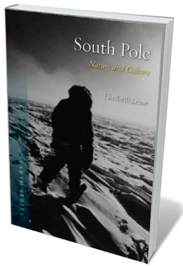
We are left with this: immersed in the unimaginable horrors perpetrated by a brutal regime, von Frisch managed to craft a hugely significant scientific discovery. Perhaps that is enough. ■

**Mark L. Winston** *is a bee biologist, professor and senior fellow at Simon Fraser University's Centre for Dialogue in Vancouver, Canada. He is the author of* Bee Time: Lessons from the Hive.
*e-mail: winston@sfu.ca*

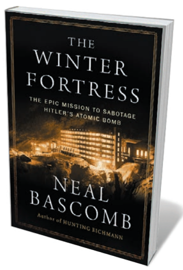
# Books in brief




**AI: Its Nature and Future**
*Margaret A. Boden* Oxford University Press *(2016)*
From search engines to satnavs, artificial intelligence (AI) permeates society. In this masterclass of a book, cognitive scientist Margaret Boden traces the evolution of AI from conceptual framing by Ada Lovelace through key research by the likes of Alan Turing and Paul Churchland, to the schism between cybernetics and symbolic computing. Traversing today's landscape, she examines the 'holy grail' of artificial general intelligence and the potential of neural networks and robots, and winnows the apocalyptic predictions from the real ethical dangers of AI misuse.

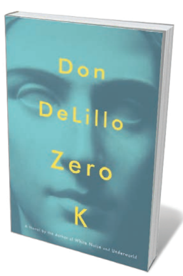



**South Pole: Nature and Culture**
*Elizabeth Leane* Reaktion *(2016)*
As the quintessence of Earthly remoteness, Antarctica has drawn hordes of scientists, iconic explorers such as Robert Falcon Scott and Roald Amundsen, and novelists who have peopled it with vast humanoid lobsters or radioactive elephant seals. Historian Elizabeth Leane tours the research, literature, exploration and geopolitical manoeuvrings that swirl around the pole. Hers is a detailed, compelling portrait of a place at once central and marginal, fantastically inhospitable and beautiful, and a mecca for physicists, government claimants and extreme tourists.




**The Winter Fortress: The Epic Mission to Sabotage Hitler's Atomic Bomb**
*Neal Bascomb* Houghton Mifflin Harcourt *(2016)*
Journalist Neal Bascomb delivers a deeply researched account of a half-forgotten episode in the Second World War: the Allied raids that sabotaged the Nazi effort to build a nuclear bomb. In 1940, the Third Reich co-opted Norway's Vemork hydroelectric plant, sole source of the heavy water ($^2H_2O$) needed for the bomb technology. Bascomb interweaves the stories of Hitler's 'Uranium Club' and of atomic chemist Leif Tronstad, who directed the Allied operation, with the thriller-esque tale of the commandos who put the plant out of action in 1943.


**Where Are the Women Architects?**
*Despina Stratigakos* Princeton University Press *(2016)*
'Male-dominated' is an understatement in architecture: in Britain alone, just 24% of architects are women, and the late Zaha Hadid was a rare star. In this slim chronicle, architectural historian Despina Stratigakos incisively catalogues the setbacks. In 1908, for instance, German architectural critic Karl Scheffler claimed that female practitioners were "irritable hermaphroditic creatures"; Ayn Rand's 1943 paean to architectural misogyny *The Fountainhead* became a university cult. Despite the equality debate, Stratigakos notes, the work of architects such as Thekla Schild remains low profile.

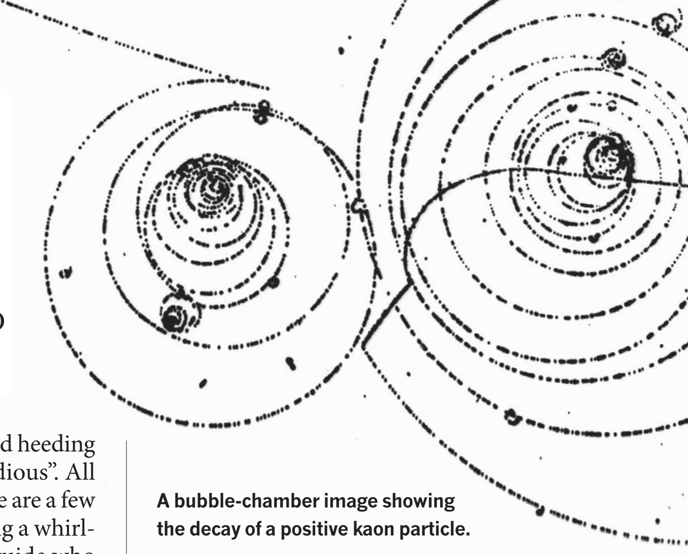



**Zero K**
*Don DeLillo* Scribner *(2016)*
Cryogenics and climate change permeate this existential science-fiction tale by novelist Don DeLillo. Set in a shadowy compound near Bishkek, Kyrgyzstan, it centres on Zero K, a "faith-based technology" that promises future immortality in cyberhuman form. Sceptical protagonist Jeff meets the cultists, views videos of catastrophes and contemplates ageing in a satirical narrative shot through with poetic lyricism. Ultimately, a celebration of life's "mingled astonishments", as a counterweight to fantasy futurism and pessimism alike. **Barbara Kiser**

PHYSICS

# Material to meaning

**Robert P. Crease** assesses Sean Carroll's attempt to construct morality out of quantum field theory.



A bubble-chamber image showing the decay of a positive kaon particle.

CERN

I don't think I have ever read anything with a bigger ambition than *The Big Picture*, physicist Sean Carroll's latest book. Physics, Carroll writes, gives us a complete picture of the foundations of nature. Although that view has had an enormous impact on cosmology, materials science and other scientific fields, its implications for meaning and morality have yet to be determined. "Our values," writes Carroll, "have not yet caught up to our best ontology." In this book, he conducts a quest to catch up.

Carroll creates his big picture as follows. Quantum field theory provides a unified perspective on the subatomic realm. Carroll calls that the "Core Theory", noting that its behaviour is fully captured by a formula called a Feynman path integral. Some features of the macro world can be directly tethered to it; others, including many concepts of thermodynamics, cannot. He calls these "emergent" features, ways of talking about the world that are not incompatible with Core Theory, yet cannot be grounded in it.

In the fun parts of *The Big Picture*, Carroll demonstrates the absurdity of adding to the Core Theory to explain the possibility of things such as an afterlife or a transcendent underlying purpose. These are easy targets. The narrative begins to get awkward when it comes to, say, conscious experiences. These, Carroll writes, are "not part of the fundamental architecture of reality"; they are emergent, a handy way of talking about what brains do. Like entropy, he argues, consciousness is a concept that "we invent to give ourselves more useful and efficient descriptions of the world". He calls his approach "poetic naturalism". By using "poetic", he means to give his blessing to ways of describing the world other than through fundamental physics — ways that, he says, can be meaningful if they are useful and don't violate the Core Theory.

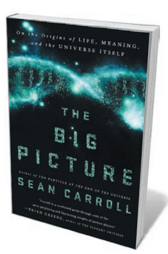
Carroll has a fluid, often engaging style, and the passages that explain science — including his appendix about the Feynman path integral — are excellent. The book brims, however, with avuncular clichés such as "Life is short, and certainty never happens". Carroll confidently defines many concepts, including belief and consciousness, as if 2,500 years of philosophy have yielded little relevant to the subject; he dismisses the task of drawing careful distinctions and heeding subtleties as "ontologically fastidious". All he finds in philosophical literature are a few interesting puzzles. It's like getting a whirlwind tour of a city from a tour guide who doesn't live there, but enthusiastically gives you capsule descriptions of favourite sites.

➲ NATURE.COM
For more on science in culture see:
nature.com/booksandarts

It is hardly surprising, therefore, that Carroll's philosophical conclusions sound profound but leave us with disappointingly empty propositions, such as, "Morality exists only insofar as we make it so, and other people might not pass judgments in the same way that we do." Outlining his own moral approach, Carroll offers a poetic naturalist's version of the Ten Commandments, the "Ten Considerations": greetings-card-like homilies such as "It Takes All Kinds".

***"Like entropy, consciousness is a concept that we invent to give ourselves more useful and efficient descriptions of the world."***

What's fascinating about *The Big Picture* is that Carroll's clarity and directness make its fundamental assumptions easy to spot, and whether you like this book will depend on whether you share them. Laboratories, as Carroll well knows, are workshops, controlled environments with unusual equipment, regulated conditions and specially trained workers. He writes from the perspective of such a worker who has come to believe that a mathematical physicist's way of thinking is just how people think — or should think — about everything, even when they are not in a workshop or when they ponder values or the existence of God. Carroll describes deciding how to be morally good, for instance, as similar to a dinner-table conversation in which, like scientists collaborating, we "talk to others about their desires and how we can work together, and reason about how to make it happen". Our group, he adds, "may include both vegetarians and omnivores, but with a good-faith effort", universal satisfaction should result.



**The Big Picture: On the Origins of Life, Meaning, and the Universe Itself**
SEAN CARROLL
*Dutton: 2016.*

Reality, too, is just what things look like from a physicist's perspective — and if it looks different to others, that is an illusion. When Carroll discusses time, he means the quantity that scientists measure. Everyday experience leads us to think that time flows in one direction, but he assures us that "in reality, both directions of time are created equal". The ontologically fastidious would say, "Not so fast!" Time as lived by humans is something else again. Both outside and even inside workshops, to be bored or expectant, to hear a melody or to plan and execute an action is not to register one moment after another, but to retain previous ones and anticipate the next in an asymmetrical flow. Determining time in the workshop is an elaborate process, and assumes that you can mark it off as you can space, and then measure the spatial movement of something, whether it is the motions of heavenly bodies in ancient times or electronic transitions in caesium atoms in ours. Yet according to Carroll, this is real time.

If we accept the strict ontology of the workshop, as Carroll does, then we get his big picture and regard lived time, conscious experience and the rest of pre-workshop life as poetic and emergent. But there are broader ontologies in which the same things — which belong to the world described by the humanities and branches of biology, for instance — are regarded as fundamental, and as the driving force for workshop activity. Carroll's is a naturalistic metaphysics.

Carroll brings tremendous passion to his writing. He is sure that honest human beings who care about the world make an effort to understand it as he does. He is right that science springs from certain basic human impulses to achieve goals and ward off threats. But where do his passion and certainty about this come from? They, too, are imported from and continue to be rooted in pre-workshop life. To find a way to talk about how scientific workshops emerge from life rather than the other way around — that would be a big picture indeed. ■

**Robert P. Crease** *is a professor in the Department of Philosophy at Stony Brook University, New York.*
*e-mail: robert.crease@stonybrook.edu*

# Correspondence

## Curb anchor scour for green shipping

In charting a course for the greening of the shipping industry (Z. Wang *et al. Nature* **530,** 275–277; 2016), we should also mitigate the scouring of seafloor biota by the massive anchors and long dragging chains dropped by a global fleet of some 68,000 ocean-going commercial vessels.

Cruise liners, too, are proliferating, with many approaching the size of supertankers. Anchoring in exotic, near-pristine locations potentially causes greater seafloor damage than it does near long-used commercial ports, which may already have been stripped by behemoths deploying anchors weighing in excess of 30 tonnes.

Ships swinging at anchor destroy seafloor animal 'forests', as well as the resources and ecosystem services they support (S. Rossi *Ocean Coast. Mgmt* **84,** 77–85; 2013). Yet the shipping industry's environmental code of practice does not recognize anchoring as a cause of concern (International Chamber of Shipping *Shipping and the Environment: A Code of Practice,* 2008).

As seaborne trade grows apace (**pictured**), there is an urgent need to assess the risks it poses to marine biodiversity. A solution could be to define safe anchorages near ports that reduce ships' physical footprints and avoid areas of high conservation value.

**Andrew R. Davis, Allison Broad** *University of Wollongong, Australia.*
*adavis@uow.edu.au*

Cargo ships off Singapore, one of the world's busiest ports.

ANEK/GETTY

## Renewables targeted before Fukushima

Masahiro Sugiyama and colleagues write that Japan expanded the role of renewables after the 2011 Fukushima Daiichi nuclear accident (*Nature* **531,** 29–31; 2016). In fact, Japan's targets for renewables were essentially unaffected by the disaster — although the country did alter its nuclear plans.

Japan's projected electricity mix for 2030 is set out in its Strategic Energy Plans. The 2014 plan (see go.nature.com/xnkn4k) aims to cut nuclear power's contribution to 20–22% by 2030, down from 53% in the 2010 plan (J. Duffield and B. Woodall *Energy Policy* **39,** 3741–3749; 2011). Fossil fuels, not renewables, are set to make up the shortfall — with the projected contribution for 2030 up by 30% compared with the 2010 plan. Meanwhile, the 2014 plan's 23% contribution from renewables by 2030 is almost unchanged (21% in the 2010 plan).

The authors rightly praise Japan's post-Fukushima attempt to expand solar power. For several decades, the country has developed this technology alongside nuclear power (R. Bointner *Energy Policy* **73,** 733–747; 2014). Japanese companies such as Sharp, Sanyo and Kyocera pioneered solar energy, whereas Hitachi, Mitsubishi and Toshiba became leaders in nuclear power. It is good news for the global climate that these technologies can be developed alongside each other.

**Aleh Cherp** *Central European University, Budapest, Hungary.*
**Jessica Jewell** *International Institute for Applied Systems Analysis, Laxenburg, Austria.*
*cherpa@ceu.edu*

## Debate over whale longevity is futile

The unquestionable importance of ethical animal husbandry aside, I doubt whether the ongoing dispute over the respective lifespans of captive and wild killer whales (*Orcinus orca*) will contribute anything to our long-term efforts to save the species (see *Nature* **531,** 426–427; 2016).

The days of keeping killer whales in captivity are in any case numbered for marine parks such as SeaWorld in the United States. And the conservation value of breeding the tiny number of captive killer whales worldwide is negligible.

In my view, we should be focusing on the real conservation plight of wild killer-whale populations around the globe (see, for example, R. Esteban *et al. Ecol. Indic.* **66,** 291–300; 2016). In the main, these are so poorly understood that entire populations are at risk of extinction (see P. J. N. de Bruyn *et al. Biol. Rev.* **88,** 62–80; 2013).

Meanwhile, we waste precious resources debating the longevity of a handful of captive animals.

**P. J. Nico de Bruyn** *University of Pretoria, South Africa.*
*pjndebruyn@zoology.up.ac.za*

## Shared goals score reproducible results

As every manager knows, the goals of the employee and the organization must be aligned for success (A. Edwards *Nature* **531,** 299–301; 2016). In my experience of industry and academic research, there is no such driver in academia.

Academics' goals are to confirm that their ideas are correct, to publish quickly and to solicit extra grant money, whereas the goal of their funding agencies is to better society. Industry and its employees have a common goal — to develop a saleable product.

This alignment means that there is little individual incentive in industry to fabricate data: drugs developed from flawed preclinical results, for example, are doomed to fail expensive multi-centre clinical trials. Irreproducibility in academic research is all too common (see *Nature* **515,** 7; 2014); in industry it is a sackable offence.

There is still some stigma attached to academics with close ties to industry, but funding agencies would do well to take note of these individuals. People in industry are not interested in working with those whose results are not reproducible.

**Eric Buenz** *Nelson Marlborough Institute of Technology, Nelson, New Zealand.*
*eric.buenz@nmit.ac.nz*

**CONTRIBUTIONS**
Correspondence may be submitted to **correspondence@nature.com** after consulting the guidelines at **http://go.nature.com/cmchno**.

For News & Views online, go to nature.com/newsandviews

MATERIALS SCIENCE

# Clockwork at the atomic scale

**Design rules for exotic materials known as polar metals have been put into practice in thin films. The findings will motivate studies of how a phenomenon called screening can be manipulated to generate new phases in metals.** SEE LETTER P.68

MARJANA LEŽAIĆ

Any science student will tell you that metals are good conductors of electric currents, whereas insulators do not allow a current to pass. A consequence of these simple facts is that a range of physical phenomena are reserved either for only the former or the latter class of material. But in this issue, Kim *et al.*[1] (page 68) show that some materials can, in certain respects, sit on both sides of the fence when fabricated according to particular design principles.

When atoms are organized into solids, they share some of their electrons. In metals, these electrons are free to move, but they remain relatively tightly bound to the atoms in semiconductors or insulators. Electrons move in accordance with electrostatic forces, which means that if a microscopic charged particle is 'dipped' into a metal, the metal's freely moving electrons become distributed so that the particle's charge does not influence electrons (or any other charged particles) a few ångströms away. This phenomenon is known as screening.

In insulators, there are no electrons that can move across the lattice and the charge distributions can be quite different from those in metals. For example, in ferroelectric materials, positively and negatively charged ions assemble as a crystal lattice and are arranged such that they form tiny electric dipoles. These dipoles interact through electrostatic forces so that they align with each other.

Ferroelectric-like phases are generally not expected to form in metals, because screening prevents the electric dipoles from 'feeling' each other's presence and so stops them from aligning. The possible existence of 'polar' metals that contain ferroelectric-like phases was nevertheless suggested more than 50 years ago[2], although few have been reported, including a handful of oxides[3,4]. Theoretical principles for designing polar metals were eventually outlined[5] in 2014. Essentially, these state that the mechanism leading to the formation of electric dipoles should be insensitive to the behaviour of the metal's conduction electrons, and they indicate ways to fulfil this requirement in crystals.

Kim and colleagues have combined experimental and theoretical efforts to achieve the first implementation of these principles in thin films. They have prepared a polar metal in thin films of neodymium nickelate ($NdNiO_3$, from the rare-earth nickelate family of compounds), which has a crystal lattice known as a distorted perovskite lattice. The authors also suggest other rare-earth nickelates that have distorted perovskite lattices as candidates for forming polar metals.

So let's talk perovskites. Simple perovskite structures can form from compounds that have the chemical formula $ABO_3$, in which $A$ and $B$ are positively charged ions (cations) and the three oxygens are negatively charged oxide ions. Ions $A$ sit at the corners of a cube, with $B$ at the centre; an oxygen ion occupies the middle of each face of the cube, forming an octahedral cage around the $B$ ion (Fig. 1a). This structural unit is repeated in all three directions to build a bulk crystal.

But in most perovskite materials, this ideal cubic structure is unstable: different cation types have different sizes, and so the oxygen cages can tilt and rotate (Fig. 1b). These distortions are often accompanied by small anti-aligned displacements of the cations from the ideal structure. However, if all of the positively charged $A$ or $B$ ions shift in the same direction with respect to the negatively charged oxygen lattice, or if both types shift, then electric dipoles form; such displacements are said to be polar.

The rotations of the oxygen cages and the polar displacements of $A$ cations ($Nd^{3+}$ ions in the case of $NdNiO_3$) compete with each other in lattices such as the $NdNiO_3$ lattice[6]: cage rotations are the preferred distortion, but if these are prevented from occurring, then polar Nd

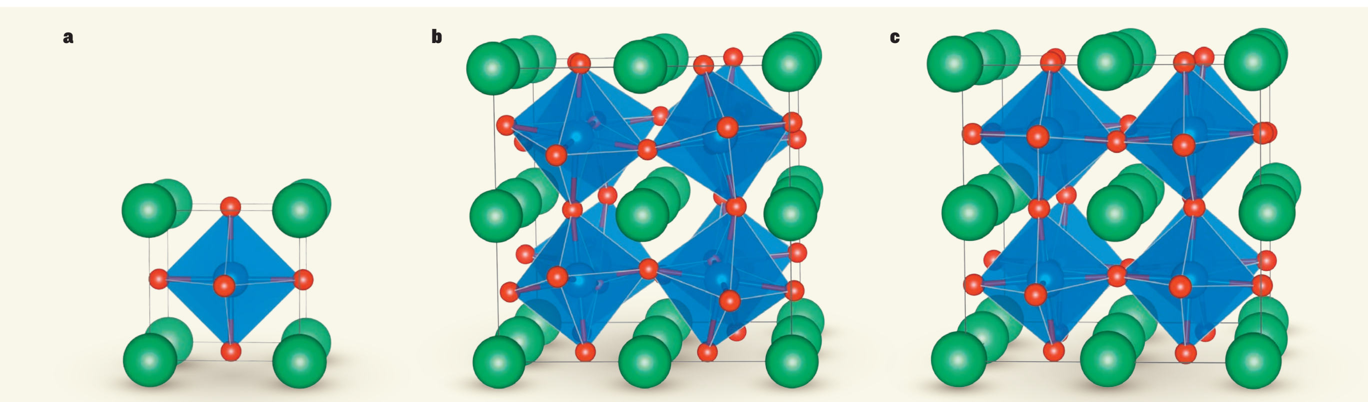



**Figure 1 | Polar metals from perovskites. a**, Perovskites are compounds that have the general formula $ABO_3$, in which $A$ and $B$ are cations and the oxygens are anions. In a simple perovskite, the oxygen anions (red) form an octahedral cage (blue) around the $B$ cation (sphere within cage), and the $A$ cations (green spheres) adopt positions at the corners of a cube around the cage. The depicted structural unit repeats in all three directions to build a bulk crystal. **b**, In most perovskites, the octahedra tilt away from the ideal positions shown in **a**, as depicted here for neodymium nickelate ($NdNiO_3$). **c**, Kim *et al.*[1] grew thin films of $NdNiO_3$ on a substrate of a different perovskite (not shown). The tilt pattern of the second perovskite counteracts that in $NdNiO_3$, and therefore reduces the tilts in the neodymium compound. This allows $NdNiO_3$ to form an exotic material known as a polar metal.

displacements take place instead. Moreover, $NdNiO_3$ is metallic at room temperature and, importantly, its conduction electrons are not derived from the $Nd^{3+}$ cation. This material therefore seems to be an ideal candidate to fulfil the design criteria for polar metals.

Kim *et al*. verified this hypothesis with first-principles calculations, which describe the behaviour of materials using the laws of quantum mechanics. The authors also theoretically determined the maximum angles of the oxygen-cage tilts that would still allow polar displacements to occur in $NdNiO_3$. The naturally occurring tilt angles are larger than the maximum angles that allow polar displacements, so the authors then developed a practical method to reduce the tilts. This involved growing thin films of the material on a substrate (the perovskite $LaAlO_3$; La is lanthanum, Al is aluminium) that was specifically chosen because its own tilt pattern counteracts that in $NdNiO_3$. The substrate can therefore rotate the cages to reduce the tilts in the neodymium compound (Fig. 1c) and so induce polar displacement of the $Nd^{3+}$ cations. This is similar to the way in which the rotating motion of cogs in some clockwork devices is transformed into linear motion elsewhere in the device — a beautiful example of atomic-scale engineering.

The authors used several experimental techniques to verify that the thin films were indeed both polar and metallic. They also showed that the crystal orientation of the substrate, which determines the 'grip' that the substrate has on the network of oxygen cages in the thin film, can tip the balance between whether an exotic polar metal or just another normal conductor forms.

The ability to engineer cage tilts at perovskite interfaces opens a fresh arena for controlling the properties of materials, but there are several notable practical aspects to consider in this approach. Interface engineering can be hampered by naturally occurring structural defects in the materials concerned, and by cation intermixing that occurs when two different materials are brought together. It is also not clear how deep into the grown film the physical properties induced by the interface can persist. Choosing a substrate that has an appropriate oxygen-cage tilt pattern is clearly crucial for fabricating materials with desired properties, but careful studies are needed to understand the parameters that control the oxygen tilts and that might make one substrate more suitable than another.

Nevertheless, the ability to use established design principles to make polar metals should reduce the scarcity of these materials, and warrants further experiments and calculations that focus on their properties and potential applications. At a fundamental level, this discovery could provide insight into the effectiveness and routes of electronic screening in complex metals. But polar metals also offer interesting functionalities — some of them are superconductors[3], for example, whereas others have strongly directionally dependent thermal properties[5]. The microscopic clockwork mechanism reported by Kim *et al*. could potentially be used to make materials that have other exotic properties, especially in 2D geometries that are suitable for integration into devices. ■

**Marjana Ležaić** *is in the Peter Grünberg Institut-1, Quantum Theory of Materials, Forschungszentrum Jülich, Jülich 52425, Germany.*
*e-mail: m.lezaic@fz-juelich.de*

1. Kim, T. H. *et al. Nature* **533,** 68–72 (2016).
2. Anderson, P. W. & Blount, E. I. *Phys. Rev. Lett.* **14,** 217–219 (1965).
3. Vaughey, J. T. *et al. Chem. Mater.* **3,** 935–940 (1991).
4. Shi, Y. *et al. Nature Mater.* **12,** 1024–1027 (2013).
5. Puggioni, D. & Rondinelli, J. M. *Nature Commun.* **5,** 3432 (2014).
6. Benedek, N. A. & Fennie, C. J. *J. Phys. Chem. C* **117,** 13339–13349 (2013).

**This article was published online on 20 April 2016.**

BIOENGINEERING

# Evolved to overcome Bt–toxin resistance

**Insects readily evolve resistance to insecticidal proteins that are introduced into genetically modified crop plants. Continuous directed evolution has now been used to engineer a toxin that overcomes insect resistance.** **SEE ARTICLE P.58**

**DANIEL DOVRAT & AMIR AHARONI**

The genetic engineering of crops to express proteins that are toxic to insects is a safe and cost-effective alternative to chemical pesticides[1]. The insecticidal toxins most commonly used in agriculture are the Cry proteins from the ubiquitous soil bacterium *Bacillus thuringiensis* (Bt). Since becoming commercially available in 1996, crops that produce Bt toxins have been widely adopted, and more than 420 million hectares have been planted around the world[2]. However, insect resistance quickly emerged as a major threat to the long-term success of such crops[2]. On page 58 of this issue, Badran *et al.*[3] present an elegant method for the continuous evolution of engineered Bt toxins, and describe a toxin that targets a new receptor on insect cells and thus overcomes existing resistance.

Bt toxins form crystalline inclusion bodies that, when ingested by insects, are solubilized and activated by gut protease enzymes[4]. The toxins then bind to specific receptors on insect midgut cells and form membrane pores that destroy the cells, killing the insect. A variety of receptors are targeted by different Bt toxins, including alkaline phosphatase, ATP-binding cassette transporters and cadherin-like proteins. The affinity and specificity of these toxin–receptor interactions underlie one of the biggest advantages of Bt toxins as pesticides: unlike broad-spectrum chemical insecticides, Bt toxins kill only specific families of insects[4], effectively suppressing pest populations without damaging their natural enemies[5] or endangering human health[1].

Alongside economic and environmental gains[6], the rapid adoption of crops engineered to produce Bt toxins has led to powerful selection pressures for resistant insects. The first field observation of substantial resistance was reported just 6 years after Bt crops were commercially introduced; since then, resistance to newly introduced toxins has appeared as little as 2 years after initial commercial availability[2]. Overall, observations accumulated over the past 20 years have repeatedly shown that insects can rapidly overcome most of the Bt-toxin crops that were designed to control them, highlighting the fierce arms race between humans and insects for crop consumption.

The evolution of insect resistance is often mediated by mutation, deletion or reduced expression of midgut-cell receptors[4]. Badran *et al.* addressed the problem of receptor-mediated resistance by engineering a widely used Bt toxin, Cry1Ac, to tightly bind to a receptor that it does not naturally target, the cadherin-like receptor from the common insect pest *Trichoplusia ni* (TnCAD). To rapidly isolate variants of Cry1Ac that have the desired characteristics, the authors used phage-assisted continuous evolution (PACE), a highly efficient method for the directed evolution of proteins.

In PACE, viruses that infect bacteria (called bacteriophage, or just phage) are made to multiply in a constant supply of host bacteria. Both the phage and the bacteria are engineered to ensure that phage infectivity depends on a specific characteristic of an evolving protein[7]. This is achieved by coupling the desired activity of the protein to the expression of a gene that is essential for phage infectivity. The target

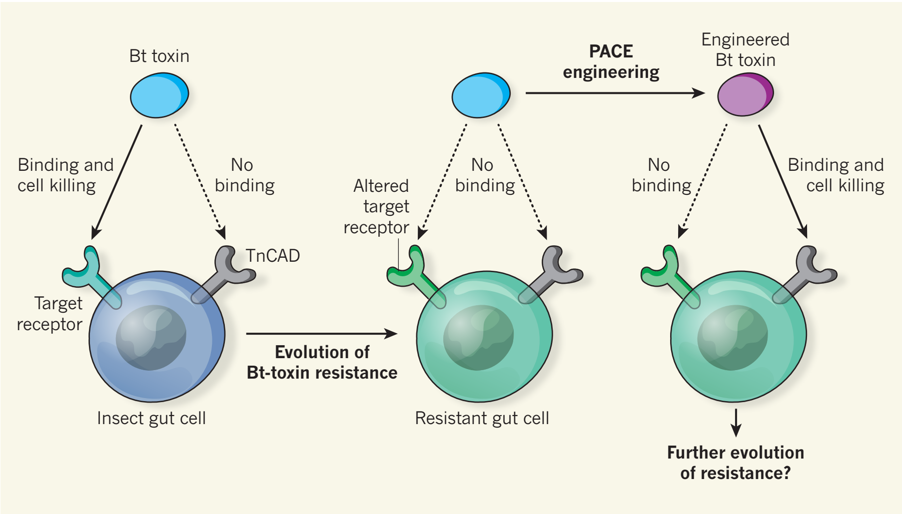



**Figure 1 | Toxin engineering by PACE.** Several proteins from the soil bacterium *Bacillus thuringiensis* (Bt) bind to receptors on insect gut cells, causing cell lysis and killing the insects. These Bt toxins have been engineered into crop plants as insecticides, but the insects rapidly evolve resistance to the toxins through mutation, deletion or downregulation of the target receptors. Badran *et al.*[3] describe an advance on the phage-assisted continuous evolution (PACE) method for directed protein evolution, in which they engineered the Bt toxin Cry1Ac such that it binds to the cadherin-like receptor of the insect *Trichoplusia ni* (TnCAD). This receptor is not bound by natural Bt toxins, so the new binding function overcomes resistance to Cry1Ac — at least, until resistant insects evolve through mutations in the TnCAD receptor.

protein for engineering, which is encoded by the phage, continuously evolves over multiple phage generations, and is under powerful selection for activity. The process is speeded up by increasing the mutation rate in the bacterial host, so that extensive genetic variability is screened in a short time.

Badran *et al.* adapted the PACE technique to evolve a tight protein–protein interaction between Cry1Ac and TnCAD. Their method (which is based on a bacterial two-hybrid system) was designed such that a stronger interaction between the evolving protein (Cry1Ac) and the binding target (a TnCAD-derived fragment) leads to increased transcription of a gene that allows for greater phage infectivity. After 22 days of continuous phage proliferation, representing more than 500 generations of replication and selection, the authors isolated multiple evolved variants of Cry1Ac. The stability of variants containing consensus mutations (mutations that appeared in several different Cry1Ac variants) was further improved by removing mutations that lead to protein destabilization. The resulting Cry1Ac variants exhibited high affinity for TnCAD, without losing their ability to bind to the native Cry1Ac receptor, and were able to efficiently kill Cry1Ac-resistant as well as susceptible insects (Fig. 1).

The incorporation of engineered Bt toxins such as these Cry1Ac variants into genetically modified crops would be a welcome addition to the limited pesticide arsenal. The evolutionary arms race will continue, of course, and it will probably be just a few years until insects evolve resistance to these new toxins as well. Nevertheless, the ability to engineer multiple toxins against target receptors of choice may prove instrumental in the future. As evermore pest species adapt to existing toxins, innovative tools and strategies to combat the evolution of resistance must be pursued to maintain the global food supply. This is especially important given the expected growth of the human population to 9.7 billion by 2050 (ref. 8), increasing the demand for crops[9].

A popular strategy to delay resistance involves 'pyramids' — crops that produce two or more toxins targeting the same pest, making the emergence of resistant insects much less likely[10]. Toxins that bind to previously untargeted insect receptors will be favourable additions to such pyramids, because they would be expected to reduce the probability of cross-resistance (when an insect that is resistant to one toxin is also resistant to another). Future work, however, may have to search for even more durable strategies, such as targeting regions on evolutionarily conserved essential receptors. ■

**Daniel Dovrat** *and* **Amir Aharoni** *are in the Department of Life Sciences and the National Institute for Biotechnology in the Negev, Ben-Gurion University of the Negev, Beersheva 84105, Israel.*
*e-mail: aaharoni@bgu.ac.il*

1. Betz, F. S., Hammond, B. G. & Fuchs, R. L. *Regul. Toxicol. Pharmacol.* **32,** 156–173 (2000).
2. Tabashnik, B. E., Brévault, T. & Carrière, Y. *Nature Biotechnol.* **31,** 510–521 (2013).
3. Badran, A. H. *et al. Nature* **533,** 58–63 (2016).
4. Pardo-López, L., Soberón, M. & Bravo, A. *FEMS Microbiol. Rev.* **37,** 3–22 (2013).
5. Lu, Y., Wu, K., Jiang, Y., Guo, Y. & Desneux, N. *Nature* **487,** 362–365 (2012).
6. Carpenter, J. E. *Nature Biotechnol.* **28,** 319–321 (2010).
7. Esvelt, K. M., Carlson, J. C. & Liu, D. R. *Nature* **472,** 499–503 (2011).
8. www.un.org/en/development/desa/news/population/2015-report.html
9. Godfray, H. C. J. *et al. Science* **327,** 812–818 (2010).
10. Carrière, Y., Fabrick, J. A. & Tabashnik, B. E. *Trends Biotechnol.* **34,** 291–302 (2016).

**This article was published online on 27 April 2016.**

PARKINSON'S DISEASE

# Guilt by genetic association

**Certain sequence variants of the α-synuclein gene are linked to the risk of Parkinson's disease. An analysis of these variants using gene-editing technology provides a possible explanation for this increased risk.** **SEE LETTER P.95**

ASA ABELIOVICH & HERVE RHINN

Genome-wide association studies have identified swathes of the human genome in which DNA sequence changes are associated with an altered likelihood that an individual will develop a given disorder, such as Parkinson's disease[1]. But the implicated DNA regions typically contain many tightly linked sequence variants that are co-inherited through the generations, and most of these are probably not involved in disease. It may therefore be impossible to identify the true culprit or culprits using genetic association studies alone. For this reason, although multiple variants in *SNCA*, the gene that encodes α-synuclein, have been associated with an increased lifetime risk of Parkinson's disease[2], the mechanisms by which they alter risk have remained enigmatic. To uncover the effects of two such *SNCA* variants, Soldner *et al.*[3] have turned to the sophisticated gene-editing technique CRISPR–Cas9. On page 95 of this issue, they report that one sequence variant increases *SNCA* expression in human neurons by reducing DNA binding

of proteins that inhibit transcription.

Single bases that vary between individuals are called single nucleotide polymorphisms (SNPs). The SNP variants in *SNCA* that have been most strongly associated with sporadic Parkinson's disease increase lifetime disease risk by around 30% (ref. 2). More than half of the world's population carries these risk-associated *SNCA* variants[4], making an understanding of their effects of paramount importance to public health.

None of the SNPs commonly associated with sporadic Parkinson's disease are predicted to alter the amino-acid sequence of the α-synuclein protein — in contrast to rare familial forms of the disease, which can be caused by changes in protein-coding regions of *SNCA*. It has therefore been proposed[5] that the common variants might instead modify gene expression. Consistent with this theory, high levels of α-synuclein accumulate in the brain tissues of people with Parkinson's disease, in abnormal neural aggregates called Lewy body inclusions that typify the disorder. Furthermore, some familial forms of the disease are caused by duplications of the entire *SNCA* gene, which leads to greatly elevated expression levels.

In pursuit of SNP variants that underlie an increased risk of Parkinson's disease, Soldner *et al.* focused on a suspect non-coding region within *SNCA*. A previous analysis[6] of human brain tissue charted molecular modifications to DNA-binding proteins that might alter gene expression and found that this region contained footprints characteristic of regulatory elements called enhancers, which influence gene expression. Soldner and colleagues investigated the region in a manner reminiscent of the precise forensic reconstruction of a crime scene, making use of CRISPR–Cas9 technology. This allows precise deletion and replacement of specific DNA sequences[7].

The investigators started with human embryonic stem cells (which can give rise to all bodily cell types) taken from an individual presumed to be unaffected by Parkinson's disease. Using CRISPR–Cas9 editing, they precisely excised a 500-base-pair stretch of DNA containing the suspect enhancer region from each of the cells' two copies of *SNCA*, which lies on chromosome 4. There are two known risk-associated SNPs in this region, called rs356168 and rs3756054. At each SNP, one variant seems to be associated with a higher risk of Parkinson's disease, whereas a different base is associated with a lower risk. Soldner *et al.* reintroduced any one of four possible SNP combinations into one of the two *SNCA* copies before inducing the human embryonic stem cells to differentiate into either neural precursors or neurons.

Next, the authors interrogated the genetically re-engineered cells using an innovative approach that precisely quantified the relative

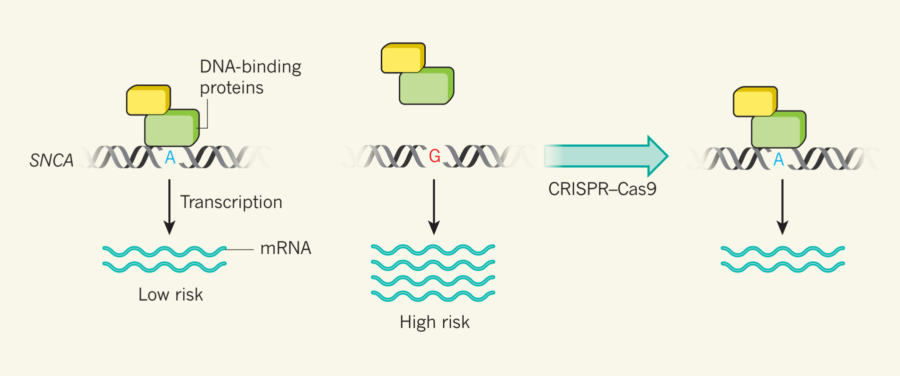



**Figure 1 | A CRISPR cross-examination.** At one nucleotide in a non-protein-coding region of *SNCA*, the gene that encodes α-synuclein, the presence of the base adenine (A) is protective against Parkinson's disease, whereas the presence of another, guanine (G), confers increased risk. Soldner *et al.*[3] report that this region regulates *SNCA* expression levels. If the two copies of the chromosome in a human cell each contain a different base at this site, gene expression is significantly higher from the risk-variant chromosome, owing in part to a reduction in the attachment of DNA-binding proteins that inhibit transcription. Using CRISPR–Cas9 gene-editing technology to remove the G and replace it with A reduces *SNCA* expression.

level of *SNCA* messenger RNA transcribed from each chromosome. The variant at rs3756054 had no effect on expression. But, remarkably, expression was 10–20% higher from chromosomes harbouring the high-risk-associated rs356168 variant than from those with the low-risk variant or those in which the enhancer was deleted (Fig. 1). Two inhibitory transcription factors, EMX2 and NKX6-1, normally bind to the DNA around this SNP, and the researchers report evidence to suggest that increased *SNCA* expression might be a direct consequence of reduced binding by these proteins to the risk variant.

Taken together, Soldner and colleagues' findings support a model whereby levels of *SNCA* expression — whether increased subtly by the presence of the high-risk variant at rs356168 or drastically, as in rare familial gene duplications — are highly correlated with the risk of Parkinson's disease. Another exciting aspect of the study is that it offers a general framework for dissecting the mechanisms underlying common disease-linked genetic variants in humans.

The work provides several avenues for further investigation. For instance, there are many SNP variants in *SNCA* that are strongly associated with Parkinson's disease but that were not interrogated in the current study. As such, Soldner *et al.* cannot rule out the possibility that the risk-associated variant at rs356168 is simply an innocent bystander. This SNP alone does not fully explain the disease risk associated with the *SNCA* region[2], and so probably has accomplices — these may have more marked effects on gene expression.

Another limitation is that Soldner and colleagues do not analyse whether their risk-associated SNPs also modulate *SNCA* expression through non-transcriptional mechanisms. For instance, disease-associated SNPs in the non-coding 3′ region of *SNCA* have been reported to regulate the processing or translation of mRNA[8]. Finally, a fundamental question is whether the SNP-dependent regulation of *SNCA* transcription seen in the authors' cell-based model is truly at work in the human brain. This could potentially be investigated by analysing brain tissue obtained at autopsy from cohorts of unaffected individuals who carry either the risk-associated or protective SNP variants.

It remains unclear how elevated levels of *SNCA* expression ultimately lead to Parkinson's disease. Nonetheless, Soldner and colleagues' findings support the pursuit of therapeutic strategies that suppress *SNCA* expression. Such efforts would complement current strategies that focus largely on improving the clearance of accumulated α-synuclein protein aggregates — for example, through the use of therapeutic antibodies. ■

**Asa Abeliovich** *and* **Herve Rhinn** *are in the Departments of Pathology, Cell Biology and Neurology, and the Taub Institute for Alzheimer's Disease and the Aging Brain, Columbia University, New York, New York 10032, USA.*
*e-mail: aa900@columbia.edu*

1. Singleton, A. B., Farrer, M. J. & Bonifati, V. *Mov. Disord.* **28,** 14–23 (2013).
2. Nalls, M. A. *et al. Nature Genet.* **46,** 989–993 (2014).
3. Soldner, F. *et al. Nature* **533,** 95–99 (2016).
4. International Parkinson Disease Genomics Consortium. *Lancet* **377,** 641–649 (2011).
5. Devine, M. J., Gwinn, K., Singleton, A. & Hardy, J. *Mov. Disord.* **26,** 2160–2168 (2011).
6. Roadmap Epigenomics Consortium. *Nature* **518,** 317–330 (2015).
7. Heidenreich, M. & Zhang, F. *Nature Rev. Neurosci.* **17,** 36–44 (2016).
8. Rhinn, H. *et al. Nature Commun.* **3,** 1084 (2012).

**The authors declare competing financial interests.**
See go.nature.com/dvrmcm for details.

**This article was published online on 20 April 2016.**

IMMUNOLOGY

# Mum's microbes boost baby's immunity

**The microorganisms that colonize pregnant mice have been shown to prime the innate immune system in newborn offspring, preparing them for life in association with microbes.**

MIHIR PENDSE & LORA V. HOOPER

Babies emerge from the womb into a world brimming with microbial life. Mammalian young inhabit a microbiologically sterile environment during fetal development, but are exposed to microbes from the moment of birth. The newborn intestine subsequently becomes colonized with trillions of microorganisms that promote digestion, block invading organisms and synthesize certain vitamins. How does the immature newborn immune system deal with this microbial onslaught? Writing in *Science*, Gomez de Agüero *et al.*[1] show that the bacteria that live in a pregnant mother's intestine provide signals that promote the development of her newborn's immune system, readying it to cope with large numbers of microbes.

Microbial colonization during the first days and weeks of newborn life has profound effects on immune-system development[2]. Many of these effects have been teased out by studies in germ-free mice, which are reared in a completely sterile setting. Germ-free mice exhibit numerous immune-system deficiencies, such as a dearth of the B and T cells that respond to foreign invaders[2]. But what happens before birth? Although the fetus lacks its own resident microorganisms, might the mother's own microbes provide cues that guide immune-system development in her offspring?

Gomez de Agüero *et al.* addressed this question using a clever experimental trick in which they exposed germ-free mice to bacteria only during pregnancy. They chose a normal bacterial resident of the gut, *Escherichia coli*, but genetically hobbled it so that it wouldn't persist in the intestine for more than a few days[3]. Pregnant mice became colonized with the hobbled strain (called *E. coli* HA107) but then returned to a germ-free state before giving birth. Thus, the developing offspring were exposed to bacteria and their products only during pregnancy — not after birth.

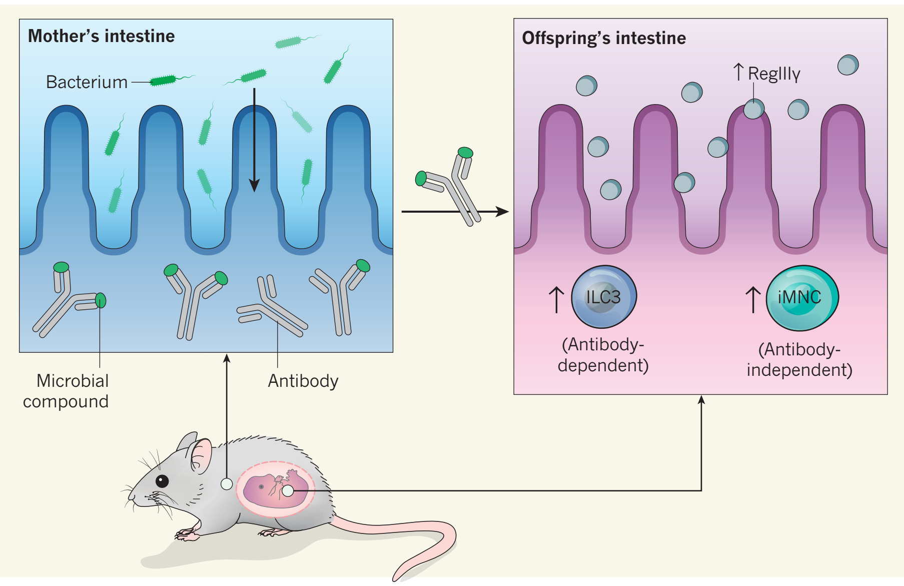



**Figure 1 | Preparation for the outside world.** Gomez de Agüero *et al.*[1] show that the presence of bacteria in the intestines of pregnant mice increases innate immunity in the offspring, and that this effect depends partly on the mother's circulating antibodies. Through an unclear mechanism, the antibodies promote transfer of microbial compounds to the developing fetus. This results in increased numbers of group 3 innate lymphoid cells (ILC3s) and increased expression of the *RegIIIγ* gene, which encodes the antimicrobial protein RegIIIγ, made by the intestinal epithelial lining. Numbers of intestinal mononuclear cells (iMNCs) are also boosted by pregnancy-specific colonization, but this increase is independent of the mother's antibodies.

The authors then studied the immune systems of offspring born to the transiently colonized mice. The newborns had increased numbers of two key immune cells that circulate throughout intestinal tissues and help to fight foreign invaders: group 3 innate lymphoid cells (ILC3s)[4] and intestinal mononuclear cells (iMNCs)[5]. Both cells are agents of the innate immune system, which is tasked with unleashing a rapid but nonspecific response to infection. Interestingly, ILC3 numbers remained elevated for several weeks after birth, suggesting that even transient colonization during pregnancy has long-term consequences for the offspring's immune system.

Although intestinal B- and T-cell numbers are boosted by colonizing germ-free mice after birth, these cells were unaffected by pregnancy-specific colonization of the germ-free mice. B and T cells are agents of the adaptive immune system, which confers long-term, specific immunity to microorganisms. Thus, pregnancy-specific colonization seems to preferentially affect cells of the innate immune system, whereas cells of the adaptive immune system are shaped largely by microbial exposure after birth.

Gomez de Agüero *et al.* found that pregnancy-specific colonization also elevates the expression of large swathes of genes in the newborn intestine. These include genes involved in metabolism, oxidative stress and innate immunity. For example, there was increased expression of the gene encoding RegIIIγ, a secreted protein that minimizes bacterial attachment to the intestinal surface[6]. These findings suggest that maternal microbes trigger a wide range of intestinal adaptations that go beyond the changes in immune-cell numbers.

How do maternal gut microbes signal to the fetus to prime development of the innate immune system? The authors first ruled out direct exposure of the fetus to live bacteria as a possible mechanism. But when they transferred serum from a mother colonized with *E. coli* HA107 into a germ-free mother, the offspring born to the serum-transplanted mice displayed the same boost in ILC3 numbers and RegIIIγ expression. Interestingly, this boost depended partly on the mother's antibodies — circulating immune molecules that bind tightly to specific antigen molecules, including those derived from bacteria. Bacterial compounds from the mother were indeed present in newborn tissues, and maternal antibodies enhanced transfer of the compounds to the offspring. It is still not clear whether this antibody-facilitated transfer is due to direct antibody binding to microbial compounds. But these findings suggest that maternal antibodies bind to microbial molecules, enter the circulation and deliver the molecules to the

developing fetus, where they prime immune-system development (Fig. 1).

When the authors investigated the chemical composition of the immunity-stimulating compounds, several were known binding partners of the aryl hydrocarbon receptor (AhR), which is essential for the development of key intestinal immune cells, including ILC3s[7]. Thus, AhR might be part of the mechanism by which maternal bacterial compounds are received by the offspring's immune system.

Do maternal microbes confer any advantages to newborns in dealing with microbial exposures? When Gomez de Agüero *et al.* exposed newborns to intestinal bacteria, those born to pregnancy-colonized mothers were better able to limit the numbers of bacteria that penetrated to deeper tissues than were those born to germ-free mothers. This suggests that the immunity boost from the mother's microbes helps to protect neonates against the pathogenic effects of bacteria, and prepares the offspring for association with large microbial communities after birth.

There are several fascinating questions that remain to be addressed. Are there other receptors besides AhR that receive maternal microbial signals in the newborn immune system? Do maternal microbial communities associated with the skin and airways also prime newborn immunity? And do maternal intestinal bacteria affect immunity in any other organs of the newborn?

A major goal in studying gut bacteria is to use their beneficial properties to improve human health. Gomez de Agüero *et al.* have laid some groundwork by identifying maternal bacterial compounds such as indole-3-carbinol — a naturally occurring ligand of AhR — that stimulate newborn immunity when fed to a pregnant mother. The work may point to new therapeutics for neonatal infectious diseases, and should encourage further investigation of how bacterial molecules augment immunity in humans. ■

**Mihir Pendse** *and* **Lora V. Hooper** *are in the Department of Immunology, University of Texas Southwestern Medical Center, Dallas, Texas 75390, USA.* **L.V.H.** *is also in the Howard Hughes Medical Institute, University of Texas Southwestern Medical Center.*
*e-mail: lora.hooper@utsouthwestern.edu*

1. Gomez de Agüero, M. *et al. Science* **351,** 1296–1302 (2016).
2. Round, J. L. & Mazmanian, S. K. *Nature Rev. Immunol.* **4,** 313–323 (2009).
3. Hapfelmeier, S. *et al. Science* **328,** 1704–1709 (2010).
4. Killig, M., Glatzer, T. & Romagnani, C. *Front. Immunol.* **5,** 142 (2014).
5. Gross, M., Salame, T. M. & Jung, S. *Front. Immunol.* **6,** 254 (2015).
6. Vaishnava, S. *et al. Science* **334,** 255–258 (2011).
7. Stockinger, B., Di Meglio, P., Gialitakis, M. & Duarte, J. H. *Annu. Rev. Immunol.* **32,** 403–432 (2014).

**This article was published online on 27 April 2016.**

CELL BIOLOGY

# Ubiquitination without E1 and E2 enzymes

**A protein in the pathogenic bacterium *Legionella pneumophila* has been found to attach the modifying molecule ubiquitin to human proteins, using a mechanism that, surprisingly, does not involve cellular E1 and E2 enzymes.** SEE LETTER P.120

SAGAR BHOGARAJU & IVAN DIKIC

Ubiquitin is a polypeptide of 76 amino acids that, when covalently attached to substrate proteins, results in either modulation of the protein's function or its destruction by the cell's proteasome machinery. Since its discovery in the late 1970s, conjugation of ubiquitin to substrate proteins has been shown to have an essential role in controlling almost all cellular processes, including cell division, DNA repair and protein synthesis[1]. The mechanism of ubiquitination is universally conserved from yeast to humans and typically proceeds through a three-enzyme cascade. Yet in this issue, Qiu *et al.*[2] (page 120) report that the bacterial protein SdeA ubiquitinates several human Rab proteins without engaging any of this cellular ubiquitination machinery.

During standard cellular ubiquitination (Fig. 1a), the ubiquitin-activating enzyme (E1) activates the carboxy terminus of ubiquitin in a process that costs one ATP molecule (the cellular energy 'currency'). The activated ubiquitin is then transferred from E1 to the ubiquitin-conjugating enzyme (E2). Finally, the ubiquitin ligase enzyme (E3) catalyses the transfer of ubiquitin from E2 to lysine amino-acid residues in the substrate protein, with or without an intermediary step of E3 self-modification[3,4].

Bacteria do not possess this ubiquitination system, but some pathogenic bacteria have evolved toxic proteins (effectors) that resemble members of the system, which they use to modulate host-cell processes to facilitate their intracellular survival and multiplication[5]. The pathogenic bacterium *Legionella pneumophila* uses about 10% of its genome (about 300 genes) to encode effectors that help it to divide and evade host-defence mechanisms[6]. Most *L. pneumophila* effector proteins have an enigmatic domain architecture that makes it difficult to predict their biochemical function on the basis of sequence similarity with other proteins, but a few effectors have been

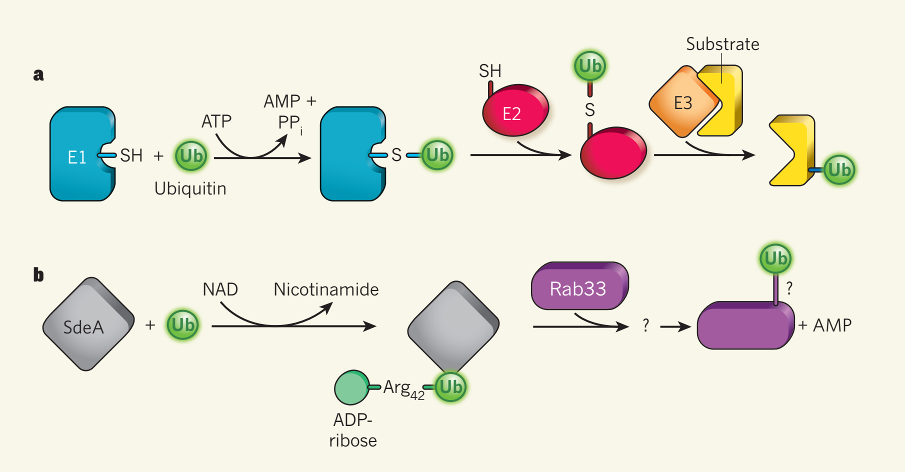



**Figure 1 | Mechanisms of ubiquitination. a,** The ubiquitination process carried out in cells from yeast to mammals involves a three-enzyme cascade. The E1 enzyme first activates the carboxy terminus of the ubiquitin molecule, using the energy from converting an ATP molecule to AMP and pyrophosphate ($PP_i$). The activated ubiquitin is attached to the sulfur of the E1 active-site cysteine residue. Ubiquitin is then transferred from E1 to E2, and E3 facilitates the transfer of ubiquitin from E2 to the substrate protein. **b,** Qiu *et al.*[2] report that the SdeA enzyme of *Legionella pneumophila* bacteria catalyses ubiquitination of the human protein Rab33 in a manner that is independent of E1 and E2. SdeA uses the cofactor NAD to add an ADP-ribose moiety to the arginine-42 ($Arg_{42}$) residue of ubiquitin in a reaction that releases nicotinamide. This is followed by modification(s) of the ADP-ribosylated ubiquitin that eventually leads to the ubiquitination of Rab33 and release of AMP, but the details of the chemistry of this transfer are not yet clear.

shown to carry out sophisticated biochemical modification of human proteins[6,7].

Effector proteins of the SidE family (SdeA, SdeB, SdeC and SidE) were previously shown to be essential for the virulence of *L. pneumophila* against its natural host amoeba[8]. By protein-sequence analysis, Qiu *et al*. found a mono-ADP ribosyltransferase (mART) motif in all members of this family. They show that the mART motif is essential for SdeA-mediated toxicity in both yeast and mammalian cell culture. Unexpectedly, however, purified SdeA exhibited no detectable ADP-ribosylation activity, indicating that it might have a different biochemical function.

To investigate further, the authors turned to Rab proteins, which are major targets of *L. pneumophila* effectors[6]. They found that co-expression of SdeA with various Rab proteins in human cells led to the covalent modification of two of these proteins, Rab1 and Rab33, which are associated with the intracellular membrane structure known as the endoplasmic reticulum. This modification depended on the mART motif of SdeA and was also seen during infection of human cells with *L. pneumophila* containing wild-type SdeA, but not when SdeA had a mutated mART motif.

Mass spectrometry revealed ubiquitin peptides in the modified Rab proteins but not in the unmodified ones, suggesting that SdeA ubiquitinates Rab proteins during *L. pneumophila* infection. However, ubiquitination of Rab33 by SdeA was not detected in an *in vitro* reaction performed in the presence of E1, ATP and various E2s, suggesting that the standard cellular enzyme cascade does not mediate this reaction. The authors then tested the ability of SdeA to modify Rab33 in the presence of both untreated and boiled human cell lysate, and observed ubiquitination in both cases, indicating that a non-protein cofactor is crucial for this process (proteins are denatured by boiling). The molecule NAD is the natural cofactor for the ADP-ribosylation mediated by other mART-containing proteins[9] — indeed, adding NAD but not ATP and/or magnesium ions (cofactors involved in standard ubiquitination) to reaction mixtures containing only SdeA, ubiquitin and Rab33 resulted in the ubiquitination of Rab33.

These observations mark the first report of substrate ubiquitination that is independent of E1 and E2 (Fig. 1b). Although the mechanistic details of SdeA-mediated ubiquitination are yet to be resolved, Qiu *et al*. present glimpses of the reaction intermediates (uncovered by mass spectrometry), which, as expected, differ from E1-dependent ubiquitination. In E1-catalysed activation, ubiquitin's carboxy terminus is modified by adenylation at the expense of an ATP; this is followed by the transfer of ubiquitin to the active-site cysteine residue of E1 and release of an AMP molecule[10]. By contrast, SdeA seems to catalyse the addition of ADP-ribose to the arginine-42 residue of ubiquitin with the help of NAD, releasing nicotinamide. The modified ubiquitin is subsequently transferred to the substrate protein through an unknown mechanism that results in the release of AMP (Fig. 1b).

In another deviation from the normal ubiquitination mechanism, SdeA shows no detectable difference in ubiquitination of Rab33 when using wild-type ubiquitin, ubiquitin lacking the two C-terminal glycine residues, or ubiquitin lacking all the surface lysine residues. It thus remains to be seen which residues of ubiquitin and Rab33 participate in the covalent linkage that is catalysed by SdeA. The authors also observed forms of Rab33 with multiple ubiquitin attachments. This may be explained by conjugation of multiple monoubiquitins or by the formation of polyubiquitin chains. Detailed structural and biochemical studies are required to address these points.

Qiu and colleagues find that SdeA-mediated ubiquitination of Rab33 has only a moderate effect on the protein's activity, and is not sufficient to explain the potent toxic effect of SdeA in cells. It is possible that more substrates exist for SdeA *in vivo*, and an unbiased screen will be needed to search for these. Undoubtedly, many researchers will also be curious about whether other proteins carry out ubiquitination independently of E1 and E2. Prime suspects for testing could be the bacterial-toxin-related mammalian proteins that contain mART motifs[11]. Qiu *et al.* have set the stage for exciting research that promises to uncover further ubiquitin chemistry with potentially far-reaching implications. ■

**Sagar Bhogaraju** *and* **Ivan Dikic** *are in the Institute of Biochemistry II, Goethe University School of Medicine, 60590 Frankfurt am Main, Germany, and in the Buchmann Institute for Molecular Life Sciences, Goethe University.*
*e-mail: ivan.dikic@biochem2.de*

1. Hershko, A., Ciechanover, A. & Varshavsky, A. *Nature Med.* **6,** 1073–1081 (2000).
2. Qiu, J. *et al. Nature* **533,** 120–124 (2016).
3. Schulman, B. A. & Harper, J. W. *Nature Rev. Mol. Cell Biol.* **10,** 319–331 (2009).
4. Joazeiro, C. A. P. & Hunter, T. *Science* **289,** 2061–2062 (2000).
5. Maculins, T., Fiskin, E., Bhogaraju, S. & Dikic, I. *Cell Res.* **26,** 499–510 (2016).
6. Hubber, A. & Roy, C. R. *Annu. Rev. Cell Dev. Biol.* **26,** 261–283 (2010).
7. Xu, L. & Luo, Z.-Q. *Microbes Infect.* **15,** 157–167 (2013).
8. Bardill, J. P., Miller, J. L. & Vogel, J. P. *Mol. Microbiol.* **56,** 90–103 (2005).
9. Barkauskaite, E., Jankevicius, G. & Ahel, I. *Mol. Cell* **58,** 935–946 (2015).
10. Haas, A. L. & Rose, I. A. *J. Biol. Chem.* **257,** 10329–10337 (1982).
11. Glowacki, G. *et al. Protein Sci.* **11,** 1657–1670 (2002).

**This article was published online on 20 April 2016.**

NUCLEAR PHYSICS

# Elusive transition spotted in thorium

**The highly precise atomic clocks used in science and technology are based on electronic transitions in atoms. The discovery of a nuclear transition in thorium-229 raises hopes of making nuclear clocks a reality.** **SEE ARTICLE P.47**

**MARIANNA SAFRONOVA**

The ability to build increasingly accurate clocks has led to technological advances such as the Global Positioning System, and has enabled tests of fundamental physics. Currently, the best clocks are based on transitions between the electronic states of atoms. On page 47 of this issue, von der Wense *et al.*[1] report the direct detection of a nuclear transition in thorium. This transition could provide the basis for a new kind of clock that would be even more precise than atomic clocks.

The first things required to build a clock[2] are some periodic events. Over the ages, the Sun rising and disappearing over the horizon provided such a reference. But to keep time accurately, one needs a periodic system that repeats its cycles at a higher frequency than this; one cycle a day will not do. Mechanical resonators, such as pendulum clocks, spring clocks and quartz-crystal resonators, were designed to serve as these periodic systems. A perfect, naturally occurring oscillator — an electronic transition in caesium atoms — eventually became the basis for the interval of time known as the second, which is defined as the duration of 9,192,631,770 cycles of this transition.

Electronic transitions in atoms, albeit with much higher frequencies than the caesium transition, now form the basis of optical atomic clocks. These contain an oscillator that is tuned to the same frequency as the chosen atomic transition, and a frequency comb (a laser-generated spectrum composed of uniformly

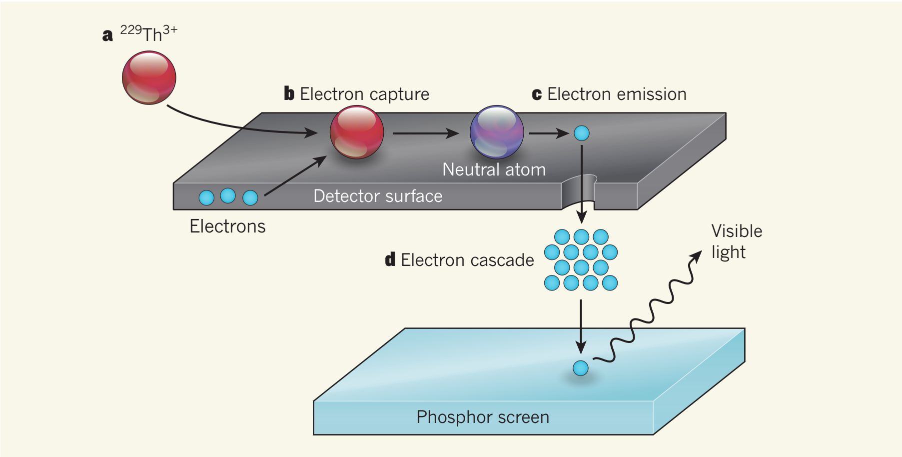



**Figure 1 | Detection of a nuclear decay process.** Von der Wense *et al.*[1] report direct evidence of an excited nuclear state of thorium-229. **a**, The authors generate thorium ions ($^{229}Th^{3+}$) in which the nucleus is in an excited state. **b**, The ions are attracted to the surface of a detector, where they capture electrons to generate neutral atoms. **c**, The excited state decays through an internal-conversion process, which causes an electron to be emitted. **d**, The emitted electron triggers a cascade of electrons, which collide with a phosphor screen, causing visible light to be produced.

spaced lines) that counts the tuned oscillation cycles. The world's most accurate atomic clock is based on an optical transition in strontium atoms trapped by laser light, and is so precise that it will neither lose nor gain 1 second in 15 billion years[3].

Von der Wense *et al.* report a milestone towards using a different type of reference oscillatory signal as the basis for a clock: a nuclear transition that occurs between an excited state (isomer) of the thorium-229 ($^{229}Th$ ) isotope and the corresponding ground state. A great attraction of a nuclear clock[4] is that it would be less sensitive to the external perturbations that cause the largest systematic errors in atomic clocks, such as electric fields and black-body radiation.

But you can't build a clock from just any nucleus. Despite the vast number of nuclear transitions that exist, almost all of them have transition frequencies that are between 10,000 and 1 million times too high for use in a nuclear clock. Only the nuclear transition in $^{229}Th$ is expected to be sufficiently long-lived and to lie in the frequency range accessible by modern laser technologies. This transition should cause the nucleus to emit radiation[5,6] that has a wavelength in the range of 150–170 nanometres, but after more than a decade of searching[4], no such emission has been observed.

Von der Wense and colleagues provide much-needed direct confirmation that the nuclear isomer responsible for the transition actually exists. Its half-life should depend on whether the thorium atom is neutral, with all its electrons present, or is one from which some electrons have been torn off, making it a positively charged ion. For neutral thorium, the isomer should decay to its nuclear ground state mainly by a process called internal conversion, which causes the emission of an electron. This decay process is fast — it is predicted[7] to have a half-life of just a few microseconds. By contrast, in thorium ions, the isomer decays by emitting an ultraviolet photon, a process that has a much longer half-life (minutes to hours, depending on the energy of the transition[7]).

The authors searched for electron emission — the signature of decay by internal conversion — in their experiments. They began by producing $^{229}Th$ ions as decay products from uranium-233. The resulting ion beam was purified to remove elements other than $^{229}Th$, and was then attracted to the surface of a detector (Fig. 1). Here, the thorium ions can acquire electrons by charge exchange with the detector's surface, forming neutral thorium atoms. The resulting thorium isomers then quickly decay by internal conversion, emitting an electron that triggers a secondary cascade of electrons. These electrons were finally accelerated to collide with a phosphor screen, generating visible light that was detected by a camera.

Von de Wense and colleagues carried out an extensive series of tests to confirm that the observed signals definitely came from the isomeric decay of $^{229}Th$, rather than any other source — in particular, short-lived nuclides formed from the $^{233}U$ decay chain, or from other isomers. In addition to providing the first direct evidence of the decay, the experiment confirmed that the transition energy is between 6.3 and 18.3 electronvolts, and that the isomeric half-life is less than 1 second for neutral thorium atoms, but more than 60 seconds for dipositive thorium ions ($Th^{2+}$).

What are the next steps towards making a nuclear clock? First, it is imperative to measure the transition energy more precisely. The smaller the uncertainty of the energy, the easier it will be to excite the transition in thorium nuclei using lasers and then to detect the ultraviolet photons that result from the isomer's decay — such detection is a crucial proof of principle before clock design can start. Second, the half-life of the isomer needs to be confirmed, to ensure that it is in the acceptable range for clock design. Finally, the energy of the $^{229}Th$ transition is much larger than those of the electronic transitions used in atomic clocks. This presents practical problems for the implementation of a nuclear clock, which will need to be overcome. If all goes well, a $^{229}Th$ clock could be about ten times more accurate than current atomic clocks[8].

But why do we want to build better clocks? Because every time that clock precision has improved, new and frequently unexpected applications have emerged. One specific reason to use $^{229}Th$ for a clock is to test for possible variation in the value of fundamental constants, such as the fine-structure constant, because the transition in the $^{229}Th$ nucleus is a particularly sensitive probe for such experiments[9]. Variation in fundamental constants has been suggested theoretically[10], and hinted at in astrophysical observations[11]. It has also been proposed[12] that ultra-precise clocks could be used to search for dark matter — the 'missing' matter in the Universe.

Moreover, networks of clocks can be used as 3D gravity sensors — the current best clocks can detect the gravitational shifts that occur when the clocks are moved to a position just 2 centimetres higher than their original position[2]. Precise and fast measurements with such a network could be used in the future to monitor volcanic magma chambers, and perhaps even to predict earthquakes[13]. In the meantime, the precision of optical atomic clocks is improving rapidly, and so the race to produce better clocks is on. What will the next decade bring? ■

**Marianna Safronova** *is in the Department of Physics and Astronomy, University of Delaware, Newark, Delaware 19716, USA.*
*e-mail: msafrono@udel.edu*

1. von der Wense, L. *et al. Nature* **533,** 47–51 (2016).
2. Ludlow, A. D., Boyd, M. M., Ye, J., Peik, E. & Schmidt, P. O. *Rev. Mod. Phys.* **87**, 637–701 (2015).
3. Nicholson, T. L. *et al. Nature Commun.* **6,** 6896 (2015).
4. Peik, E. & Okhapkin, M. *C. R. Phys.* **16,** 516–523 (2015).
5. Beck, B. R. *et al. Phys. Rev. Lett.* **98,** 142501 (2007).
6. Beck, B. R. *et al. Proc. 12th Int. Conf. Nuclear Reaction Mechanisms* LLNL-PROC-415170 (2009).
7. Tkalya, E. V., Schneider, C., Jeet, J. & Hudson, E. R. *Phys. Rev. C* **92,** 054324 (2015).
8. Campbell, C. J. *et al. Phys. Rev. Lett.* **108,** 120802 (2012).
9. Berengut, J. C., Dzuba, V. A., Flambaum, V. V. & Porsev, S. G. *Phys. Rev. Lett.* **102,** 210801 (2009).
10. Uzan, J.-P. *Living Rev. Relativ.* **14,** 2 (2011).
11. Webb, J. K. *et al. Phys. Rev. Lett.* **107,** 191101 (2011).
12. Derevianko, A. & Pospelov, M. *Nature Phys.* **10,** 933–936 (2014).
13. Bondarescu, M., Bondarescu, R., Jetzer, P. & Lundgren, A. *EPJ Web Conf.* **95,** 04009 (2015).

doi:10.1038/nature17669

# Direct detection of the $^{229}$Th nuclear clock transition

Lars von der Wense[1], Benedict Seiferle[1], Mustapha Laatiaoui[2,3], Jürgen B. Neumayr[1], Hans-Jörg Maier[1], Hans-Friedrich Wirth[1], Christoph Mokry[3,4], Jörg Runke[2,4], Klaus Eberhardt[3,4], Christoph E. Düllmann[2,3,4], Norbert G. Trautmann[4] & Peter G. Thirolf[1]

**Today's most precise time and frequency measurements are performed with optical atomic clocks. However, it has been proposed that they could potentially be outperformed by a nuclear clock, which employs a nuclear transition instead of an atomic shell transition. There is only one known nuclear state that could serve as a nuclear clock using currently available technology, namely, the isomeric first excited state of $^{229}$Th (denoted $^{229m}$Th). Here we report the direct detection of this nuclear state, which is further confirmation of the existence of the isomer and lays the foundation for precise studies of its decay parameters. On the basis of this direct detection, the isomeric energy is constrained to between 6.3 and 18.3 electronvolts, and the half-life is found to be longer than 60 seconds for $^{229m}Th^{2+}$. More precise determinations appear to be within reach, and would pave the way to the development of a nuclear frequency standard.**

The first excited nuclear state of $^{229}$Th is one of the most exotic states in the whole nuclear landscape: of the known 176,000 nuclear levels[1], it possesses the lowest excitation energy, about 7.8 eV (refs 2, 3). Although there is one other nuclear excitation known[1] to have a transition energy below 1 keV ($^{235m}$U, 76 eV), typical nuclear excitation energies are $10^4$ to $10^6$ times larger[4] (Fig. 1).

The $^{229}$Th nucleus was first considered in 1976 to possess an isomer (a metastable nuclear state) with an excitation energy below 100 eV (ref. 5). Further measurements supported its existence[6,7] and stepwise improvement in techniques led in 1994 to a measured value of the excitation energy of $3.5 \pm 1.0$ eV (ref. 8). However, in 2007 a microcalorimetric measurement suggested a value of $7.8 \pm 0.5$ eV, corresponding to a wavelength near 160 nm for radiation emitted in the decay to the ground state[2,3]. This uniquely low nuclear transition energy can potentially bridge the fields of nuclear and atomic physics, as it conceptually allows for optical laser excitation of a nuclear transition[9]. This in turn has stimulated thoughts about transferring existing knowledge of laser manipulation of the electronic shell to a nuclear system, leading to interesting proposed applications such as a nuclear laser[10], nuclear quantum optics[11], and a nuclear clock[12,13].

Besides the low excitation energy $E$, a radiative isomeric half-life in the range of minutes to hours has been predicted[14–16], resulting in a relative linewidth as low as $\Delta E/E \approx 10^{-20}$. These unique features render this transition an ideal candidate for a nuclear clock[12], which may outperform existing atomic-clock technology owing to potentially improved compactness and expectedly higher resilience against external influences[13,17]. Two ways to establish a nuclear clock are currently being investigated; one based on $^{229}Th^{3+}$ stored in a Paul trap[13,18,19], and the other based on $^{229}$Th embedded in a crystal-lattice environment[12,20–23].

The immediate impact and far reaching implications of a nuclear clock become clear when considering current applications of existing atomic-clock technology[24]. Moreover, a nuclear clock promises intriguing applications in fundamental physics—for example, the investigation of possible time variations of fundamental constants[25–28].

To date, experimental knowledge of the isomer has been inferred indirectly[2,3,5–8]. However, a direct detection was still pending. Such a direct detection would not only give further evidence for the isomer's existence, but also pave the way to precise studies of the half-life, excitation energy and decay mechanism of the isomeric state, which are the basis for a direct optical excitation[29]. This has motivated significant experimental effort aimed at further validation of the isomer's existence[30–32] and direct detection of the isomeric de-excitation[21,33–38]. For a detailed overview, we refer the reader to a recent review and references therein[39]. Despite decade-long efforts, none of these previous attempts has conclusively reported the isomer's direct detection. Here we report the direct observation of this elusive isomeric decay. This direct detection paves the way to the precise determination of all decay parameters relevant for optical excitation.

## Experimental setup

Decay of the $^{229}$Th isomeric state of the neutral thorium atom occurs predominantly by internal conversion (IC) with emission of an electron[15,16], which is used as a key signature for identifying the $^{229}$Th isomer (spin and parity $3/2^+$, Nilsson quantum numbers [631]) to ground state ($5/2^+$ [633]) de-excitation. A short half-life in the microsecond range was predicted for this case[15,16]. This is because the 6.31 eV first ionization potential of thorium is below the suggested energy of the isomeric transition. In a higher charge state (that is, thorium ions), the IC process is energetically forbidden and radiative decay may dominate. In this case, the half-life is expected to increase significantly to minutes or hours. Searches for an IC decay with a half-life of a few milliseconds or longer for neutral thorium have already been conducted[40]. Our experimental setup[41], as shown in Fig. 2, was designed for the detection of a low-energy IC decay of shorter half-life. A schematic of the experimental process is shown in Extended Data Fig. 1.

The isomeric state in $^{229}$Th can be populated via a 2% decay branch in the $\alpha$ decay of $^{233}$U (ref. 42). For detection of the isomer, a $^{233}$U source is placed in a buffer-gas stopping cell[43] (Extended Data Fig. 2) into which $^{229}$Th ions, produced in the $\alpha$ decay of $^{233}$U, are recoiling, along with $^{229}$Th daughter products if present in the source. These $\alpha$-recoil ions are stopped in 40 mbar of ultra-pure helium. Removing the up to 84 keV kinetic recoil energy (significantly greater than the isomer energy of a few electronvolts) is essential for the experiment. During the stopping process charge exchange occurs, producing predominantly thorium in the 2+ and 3+ charge states. These ions are guided by an

[1]Ludwig-Maximilians-Universität München, 85748 Garching, Germany. [2]GSI Helmholtzzentrum für Schwerionenforschung GmbH, 64291 Darmstadt, Germany. [3]Helmholtz-Institut Mainz, 55099 Mainz, Germany. [4]Johannes Gutenberg-Universität Mainz, 55099 Mainz, Germany.

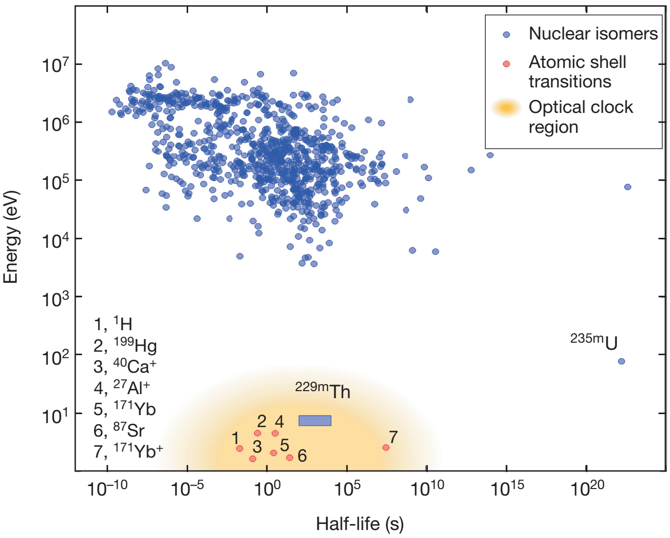



**Figure 1 | Energy–half-life distribution.** Blue circles, known nuclear isomeric states[4] (darkening occurs where circles overlap); red circles (numbered, key at bottom left), selected atomic shell transitions used for frequency metrology. Orange region, the parameter space currently accessible for optical clocks. $^{229m}$Th (expected region shown as a blue box) exhibits a uniquely low excitation energy, and is the only known promising isomer for the development of a nuclear-based frequency standard using existing technology. One other nuclear isomer with an energy below $10^3$ eV is known ($^{235m}$U, bottom right), however, it has a significantly longer half-life. Purely radiative half-lives are shown for $^{229m}$Th and $^{235m}$U, this being the relevant parameter for the development of a nuclear clock.

electric field through a radio-frequency (RF) and direct-current (DC) funnel system towards the buffer-gas stopping cell exit, where they are extracted by a supersonic Laval nozzle and injected into a radio-frequency quadrupole (RFQ) structure. While the ions are guided by the electric fields provided by the RFQ, the remaining ambient helium gas pressure leads to phase-space cooling, such that a recoil-ion beam with submillimetre diameter is formed at the RFQ exit. There, most of the daughter nuclides from the $^{233}$U decay chain are still present, some of which are short-lived $\alpha$ or $\beta^-$ emitters. A quadrupole mass-separator (QMS) is used for ion-beam purification, such that only $^{229}$Th remains. Subsequently, the thorium ions are guided with the help of a triodic guidance structure with a 2-mm-diameter orifice towards a microchannel plate (MCP) detector, used for low-energy electron detection. The ions are collected in a soft landing at low kinetic energy (50–75 eV, depending on the charge state) directly on the MCP detector (operated at −25 V surface voltage), which is placed in front of a phosphor screen. The latter is monitored by a charge-coupled device (CCD) camera, allowing for a spatially resolved signal detection.

## Isomer detection

Because stopping and extraction of $^{229}$Th occurs in the form of ions and takes only a few milliseconds, there is no significant isomer decay during the time of flight. However, when the ions come into contact with the MCP surface, charge exchange occurs forming neutral thorium atoms, for which rapid IC is expected to dominate the decay of the isomeric state. This process releases a conversion electron, which is accelerated into a microchannel of the MCP detector, triggering the emission of secondary electrons. The electron 'cloud' thus produced is accelerated towards the phosphor screen, where the electronic-impact signal is converted into visible light that is detected with the CCD camera. This detection technique has some similarity to the MCP-based detection of metastable molecular states in chemistry[44], and has already been successfully applied to the detection of $^{235m}$U (ref. 40). A schematic drawing of the detection process on the MCP surface from a microscopic perspective is shown in Fig. 3 (see also Methods section).

The $^{233}$U source (denoted below as source 1) consists of a layer of $^{233}UF_4$ (of activity level ~200 kBq) that was evaporated onto a 20-mm-diameter stainless steel plate. A complete mass scan of ions extracted from this source is shown in Fig. 4a. We measured the $^{229}Th^{3+}$ ion extraction rate from source 1 to be about $10^3\,s^{-1}$ (ref. 45). Assuming that 2% of the ions are in the isomeric state[42] and also accounting for an MCP detection efficiency for low-energy electrons of about 1.5% (ref. 46), a count rate of ~0.3 counts $s^{-1}$ is expected. The isomeric-decay signal obtained when extracting $^{229}Th^{3+}$ for 2,000 s is shown in Fig. 4c. Signals were acquired within a centred field of view as obtained within a 20-mm-diameter aperture (see Methods for details of image readout). The spatially integrated decay count rate is $0.25 \pm 0.10$ counts $s^{-1}$ and in good agreement with the expectations. The error was estimated to also account for changes in the $^{229}Th^{3+}$ extraction efficiency. The MCP exhibits a low dark count rate of 0.01 counts $s^{-1}\,mm^{-2}$, leading to a signal to background ratio of about 8:1. An overview of different measurements performed under the same conditions is shown in Fig. 4b. Each row corresponds to an individual uranium source, as will be detailed in the following section, while each column corresponds to a different extracted ion species, as indicated by the arrows from the mass scan. Clear signals are seen when extracting $^{229}Th^{2+}$ and $^{229}Th^{3+}$ (Fig. 4b, first row). For completeness, measurements were also performed while extracting $^{229}Th^{1+}$. In this case, no signal could be obtained, which might be attributable to the very low extraction efficiency of just 0.3% for $Th^{1+}$, compared to 5.5% for $Th^{2+}$ and 10% for $Th^{3+}$ (ref. 45).

## Signal identification

In order to prove that the detected signal originates from the $^{229}$Th isomeric decay, comparative measurements were performed which allowed us to exclude all potential background sources. These can be grouped into four categories: (A) background attributed to the kinetic energy or charge state of the impinging ions, (B) background signals from setup components ($^{233}$U source, buffer-gas stopping and extraction,

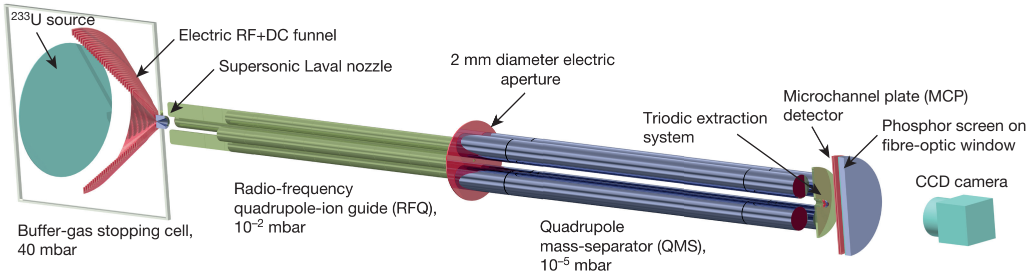



**Figure 2 | Schematic of the experimental setup.** The $^{233}$U source is mounted in front of an RF+ DC funnel placed in a buffer-gas stopping cell[43]. $^{229}$Th $\alpha$-recoil ions, emitted from the source, are extracted for ion-beam production in a radio-frequency quadrupole system (RFQ). After mass purification of the ion beam with the help of a quadrupole mass-separator (QMS), the ions are attracted at low kinetic energy (for soft landing) onto the surface of a microchannel plate detector (MCP). There the $^{229}$Th isomeric decay signals are detected (for details, see text and Methods).

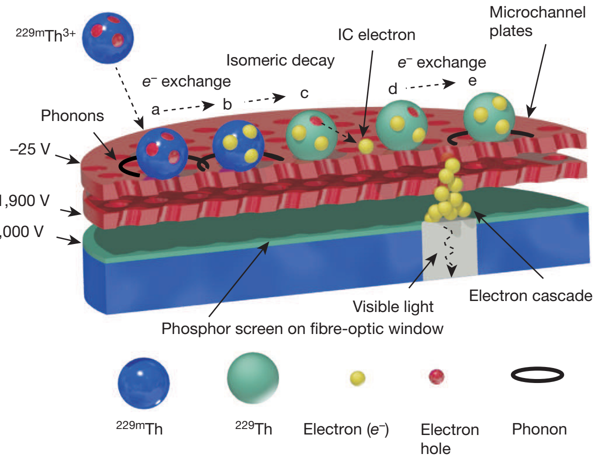



**Figure 3 | Schematic drawing of the isomer detection process.** The process, shown broken into steps a–e on the microchannel surface, is as follows. Step a, a $^{229m}Th^{3+}$ ion impinges on the MCP surface. The thorium ion in the isomeric state is visualized as a blue sphere. Step b, electron capture on the surface. The energy is dissipated in form of phonons (indicated as black circles). Electrons are visualized as yellow spheres. Step c, an IC electron is released by the isomeric decay. Step d, the IC electron triggers a secondary-electron cascade, which is accelerated towards the phosphor screen. Step e, the hole, left by the IC process, is filled by electron attachment on the MCP surface. Again, phonons are produced.

QMS, MCP detection system), (C) signals originating from the thorium atomic shell (long-lived excited states or chemical reactions on the MCP surface) and (D) signals caused by short-lived nuclides or other isomers (not of $^{229}$Th). Most of the possible background effects were excluded in several ways. An overview is shown in Extended Data Table 1.

Ionic energy, as carried in the form of momentum or ion charge state, may lead to the release of electrons on the MCP surface. In order to exclude this type of background (type A), a $^{233}U^{2+}$ mass peak (originating from sputtering of the source), which has an intensity similar to that of the $^{229}Th^{2+}$ mass peak (Fig. 4a), is used for comparison (Extended Data Table 1 no. 1). During 2,000 s of continuous extraction of $^{233}U^{2+}$, no MCP signal was obtained (Fig. 4b, first row).

Furthermore, a measurement of the signal intensity as a function of the MCP surface voltage was carried out for $^{229}Th^{2+}$ and $^{233}U^{2+}$ (Extended Data Table 1 no. 2, Fig. 5a). For this purpose, each isotope was extracted for 1,200 s for every data point. For MCP surface voltages between −100 V and −40 V, the remaining ion-impact signal decreases as the kinetic energy of the ions is reduced. While the uranium signal is effectively reduced to zero, a thorium signal remains. A sharp cut-off of this signal occurs at zero kinetic energy, when the ions can no longer approach the MCP surface. An enhancement of the signal intensity is observed just before the cut-off, and is attributed to IC electrons back-attracted into the MCP surface. The absence of a similar sharp cut-off for uranium clearly excludes any cause of the signal by ion impact or charge state. Further, these measurements also exclude all potential background caused by the setup components (type B), which would be constant throughout the measurements.

Thorium atomic shell effects, such as a long-lived atomic excitation or a chemical reaction between thorium and the MCP surface, could potentially contribute background (type C). To exclude this possibility it is sufficient to perform a comparative measurement with $^{230}$Th where such effects would be identical (Extended Data Table 1 no. 3). For this purpose, a $^{234}$U source was employed (270 kBq, electrodeposited onto a titanium sputtered silicon wafer, denoted below as source 2). The $^{230}$Th $\alpha$-recoil ions emerging from this source were accumulated on the surface of the MCP detector for 2,000 s, just as for $^{229}$Th. For $^{230}$Th, however, no signal is detected (Fig. 4b, second row), which proves that the signal obtained for $^{229}$Th cannot be caused by an atomic shell effect. This measurement also provides further exclusion of background of types A and B. In this way most of the systematic background effects are excluded.

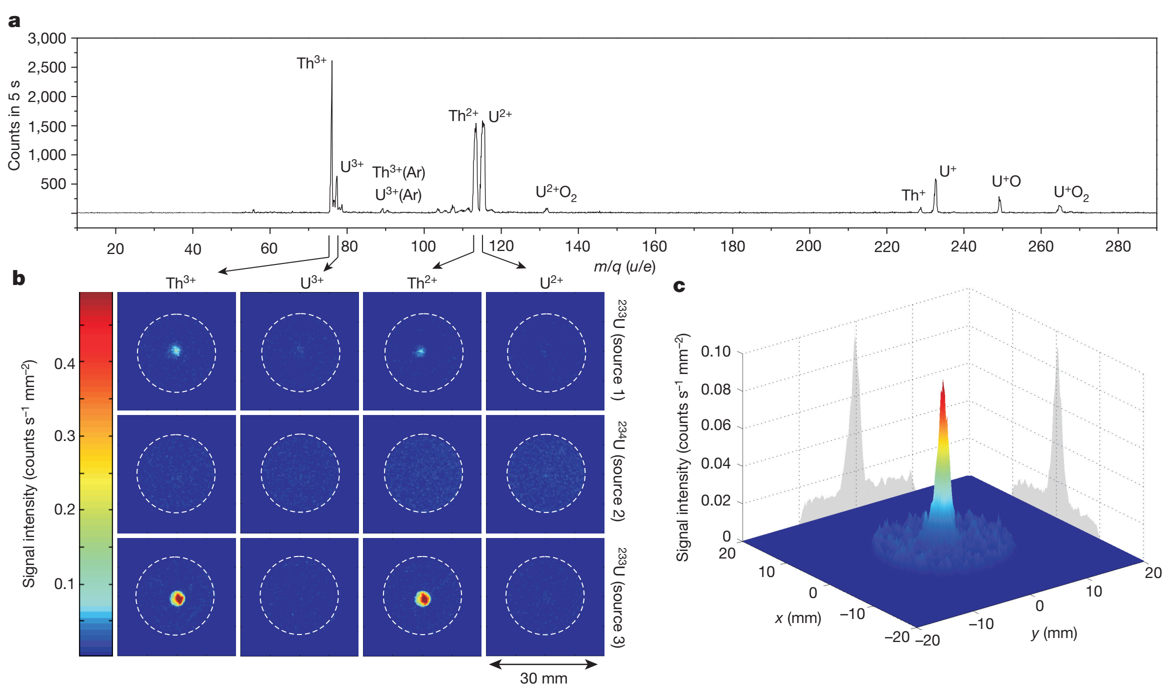



**Figure 4 | Signal comparison. a**, Complete mass scan performed with the $^{233}$U source 1 (ref. 45). Units are given as atomic mass ($u$) over electric charge ($e$). **b**, Comparison of MCP signals obtained during accumulation of thorium and uranium in the 2+ and 3+ charge states (see individual extracted ions at top, arrowed from mass scan); $^{233}$U and $^{234}$U sources were used (the source number is given on the right-hand side of each row). Each image corresponds to an individual measurement of 2,000 s integration time (20 mm diameter aperture indicated by dashed circles). Measurements were performed at about −25 V MCP surface voltage in order to guarantee soft landing of the ions. **c**, Signal of the $^{229}$Th isomeric decay obtained during $^{229}Th^{3+}$ extraction with source 1. A signal area diameter of about 2 mm (full-width at half-maximum) is achieved. The obtained maximum signal intensity is 0.08 counts s$^{-1}$ mm$^{-2}$ at a background rate of about 0.01 counts s$^{-1}$ mm$^{-2}$.

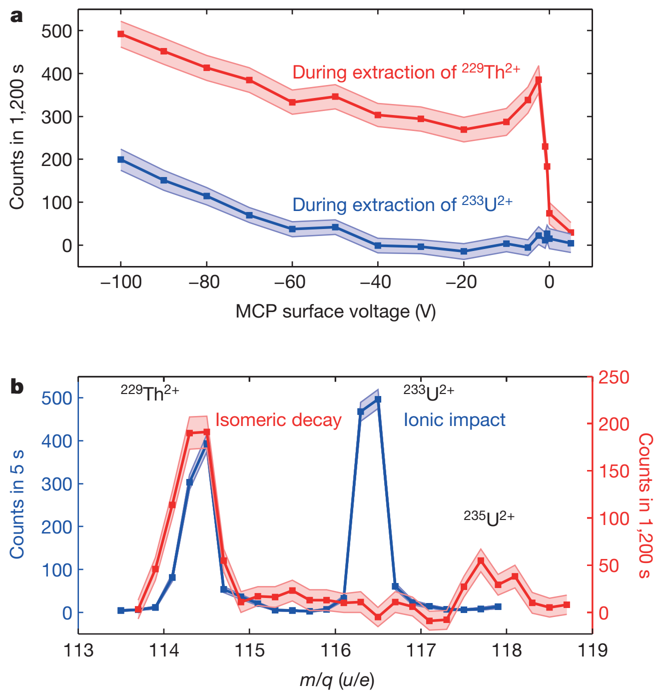



**Figure 5 | Background corrected $^{229}$Th isomeric decay signals. a,** $^{229}Th^{2+}$ signal (red) compared to $^{233}U^{2+}$ (blue) as a function of the MCP surface voltage. Errors are indicated by shaded bands, given as the estimated s.d. of the Poisson distribution for sample size $n=1$ ($\sqrt{S+N}$, where $S$ and $N$ denote the total and the background count numbers of about 60 counts, respectively). **b,** Signal of extracted ions as a function of the mass-to-charge ratio behind the QMS for MCP surface voltages of $-25$ V (isomeric decay, red) and $-2{,}000$ V (ion impact, blue). Note the different integration times and axis scales. Besides the signal at 114.5 $u/e$ (corresponding to $^{229}Th^{2+}$), a further signal at 117.5 $u/e$ occurs, which originates from the isomeric decay of $^{235}$U ($^{239}$Pu was shown to be contained in the source material by $\alpha$ spectroscopy[45], the isomer is populated by a 70% decay branch and the extraction rate is too small to be visible in the ion-impact signal).

In earlier experiments, direct identification of the $^{229m}$Th isomeric decay was prevented in part by radioactive decay of short-lived daughter nuclides[39]. Our experiments focused specially on this type of potential background (type D), which we have been able to exclude in four independent ways. A QMS is used for the extraction of ions with a well-defined mass-to-charge ratio from the buffer-gas stopping cell. The achieved mass-resolving power of $m/\Delta m=150$ is sufficient for the complete separation of the $\alpha$-recoil ions with a difference of four or more atomic mass units (Extended Data Fig. 3). Figure 5b shows the signal intensity as a function of the selected mass-to-charge ratio $m/q$ for MCP surface voltages of $-25$ V and $-2{,}000$ V. At a $-2{,}000$ V surface voltage (blue), the ion-impact signal is observable and the $^{233}U^{2+}$ and $^{229}Th^{2+}$ mass peaks are of comparable amplitude. At the $-25$ V surface voltage (red), the $^{233}U^{2+}$ mass peak completely vanishes, since no ion-impact signal is detected. $^{229}Th^{2+}$, in contrast, reveals a remaining component, which is clearly restricted to the $^{229}Th^{2+}$ mass peak. However, molecular sidebands may be populated by nuclides of lower masses (for example, $^{213}Bi^{16}O$ has the same mass as $^{229}$Th and is a $\beta^-$ emitter in the $^{233}$U decay chain with a 45.6 min half-life). Thus restriction to the $m/q$ value of $^{229}Th^{2+}$ does not exclude short-lived daughter nuclides as signal contributions.

The first way to exclude this sort of background is obtained from the parallel observation of the signals in the 2+ and the 3+ charge states (Extended Data Table 1 no. 4, Fig. 4b, first row), because only thorium is extracted to a significant extent in the 3+ charge state because of its low third ionization potential[45] (see Extended Data Table 2 for comparison). Experimentally, a suppression of three to four orders of magnitude for the short-lived daughters in the 3+ charge state compared to the 2+ charge state was obtained[45].

The second way to exclude nuclear background is based on a comparison (Extended Data Table 1 no. 5) which was performed with a newly available chemically purified $^{233}$U source (source 3, 290 kBq, same geometry as source 2). The factor of chemical purification of the short-lived daughter nuclides was measured to be $\geq 250$; if signals were originating from nuclear background, a drastically reduced signal intensity should occur when this new source was used. This reduction is, however, not observed, and instead the signal increases by a factor of $\sim 13.5$ owing to a larger $^{233}$U content and a reduced source thickness, leading to a higher $\alpha$-recoil efficiency (Fig. 4b, third row).

The third and fourth ways of excluding nuclear background are discussed in the Methods section. Consequently, the nuclear isomeric transition in $^{229}$Th is the only possible explanation for the observed signal.

## Half-life and energy constraints

Direct detection of the $^{229m}$Th isomeric-decay signal provides constraints on the half-life of the isomer, which is found to be heavily charge-state dependent. Two different measurements were performed (see Methods for details): the first to estimate an upper limit for the isomeric half-life in the neutral thorium atom, and the second to infer a lower limit for the isomer's lifetime in $^{229}Th^{2+}$. These measurements allow us to draw conclusions about the isomeric energy, as the half-life changes depending on whether the IC decay channel is energetically permitted or not.

For the neutral thorium atom, $^{229m}$Th is predicted to decay predominantly by IC with a half-life as short as microseconds[15,16]. Experimentally, an upper limit for the isomeric half-life in neutral thorium was found by $^{229(m)}Th^{2+}$ ion-beam pulsing (brackets are used for the mixed-nuclide ion beam). Images were acquired directly after the ion-pulse had struck the MCP surface, leading to the formation of neutral thorium by charge exchange. In this way the half-life was determined to be less than one second, confirming that the isomeric IC decay-channel is energetically allowed. This in turn gives a strong indication that the isomeric energy is above the first ionization potential of thorium, 6.31 eV.

An isomeric half-life of minutes to hours has been predicted for $^{229m}$Th in a charge state $\geq 1+$, where IC is energetically forbidden[15,16]. In order to confirm this prediction, $^{229(m)}Th^{2+}$ ions were stored in the RFQ before acquiring the isomeric-decay signal. The half-life range probed in this way was limited by the maximum ion storage time in the RFQ, which is about 60 s. Still, after this time, significant isomeric decay was detected, suggesting the isomeric lifetime in $Th^{2+}$ to be longer than 60 s. This long half-life can only be explained if the isomeric IC decay-channel is energetically forbidden for $^{229}Th^{2+}$. Thus the isomeric energy must be below the third ionization potential of thorium, 18.3 eV.

On the basis of the half-life estimates, the value of the isomeric energy is deduced to be between 6.3 and 18.3 eV (that is, between the first and third ionization potential of thorium). This energy range is consistent with today's most accepted value[2] and promising for the development of a nuclear clock based on thorium ions.

## Discussion and perspectives

The efficient production of a low-energy, highly pure $^{229(m)}$Th ion beam has enabled the successful direct observation of the decay of the $^{229}$Th isomer to its ground state, using a spatially decoupled isomer population and isomeric decay combined with efficient mass separation using a QMS.

This measurement is not only a further proof of the isomer's existence, which has been controversial[29,47], but also provides a detection method that could be used as a tool to probe different processes for isomer population—for example, via direct laser excitation[19] or electronic bridge processes[39]. Further, in the nuclear-clock concept, the observed IC decay could be used to probe the isomeric population

to provide an alternative to the proposed double-resonance method[12]. Most importantly, this direct detection paves the way to precise determination of the isomer's decay parameters. The isomeric half-life could be probed by applying a cryogenically cooled Paul trap[48], which allows longer ionic storage times. A more precise energy value could be determined by applying a hemispherical electron energy analyser[49] with an energy resolution of a few millielectronvolts (see Methods for details). This would allow the possibility of developing a laser system that could ultimately bring all-optical control of this nuclear transition and thus provide a template for coherent manipulation of nuclei in general[50]. The construction of a nuclear frequency standard based on this $^{229}$Th isomeric transition would open new perspectives in ultra-precise frequency metrology that are expected to have implications for both technology and fundamental physics.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 16 December 2015; accepted 16 March 2016.**


1. NNDC interactive chart of nuclides. http://www.nndc.bnl.gov/chart (accessed 16 June 2015).
2. Beck, B. R. *et al.* Energy splitting of the ground-state doublet in the nucleus $^{229}$Th. *Phys. Rev. Lett.* **98,** 142501 (2007).
3. Beck, B. R. *et al. Improved Value for the Energy Splitting of the Ground-state Doublet in the Nucleus $^{229m}$Th*. LLNL-PROC-415170 (Lawrence Livermore National Laboratory, 2009).
4. Tuli, J. K. *Nuclear Wallet Cards* 8th edn (National Nuclear Data Center, Brookhaven National Laboratory, 2011).
5. Kroger, L. A. & Reich, C. W. Features of the low energy level scheme of $^{229}$Th as observed in the $\alpha$-decay of $^{233}$U. *Nucl. Phys. A* **259,** 29–60 (1976).
6. Burke, D. G., Garrett, P. E., Qu, T. & Naumann, R. A. Additional evidence for the proposed excited state at $\leq$ 5 eV in $^{229}$Th. *Phys. Rev. C* **42,** 499–501 (1990).
7. Burke, D. G., Garrett, P. E., Qu, T. & Naumann, R. A. Nuclear structure of $^{229,231}$Th studied with the $^{230,232}$Th(d,t) reactions. *Nucl. Phys. A* **809,** 129–170 (2008).
8. Helmer, R. G. & Reich, C. W. An excited state of $^{229}$Th at 3.5 eV. *Phys. Rev. C* **49,** 1845–1858 (1994).
9. Tkalya, E. V. Properties of the optical transition in the $^{229}$Th nucleus. *Phys. Usp.* **46,** 315–320 (2003).
10. Tkalya, E. V. Proposal for a nuclear gamma-ray laser of optical range. *Phys. Rev. Lett. A* **106,** 162501 (2011).
11. Bürvenich, T. J., Evers, J. & Keitel, C. H. Nuclear quantum optics with X-ray laser pulses. *Phys. Rev. Lett.* **96,** 142501 (2006).
12. Peik, E. & Tamm, C. Nuclear laser spectroscopy of the 3.5 eV transition in $^{229}$Th. *Europhys. Lett.* **61,** 181–186 (2003).
13. Campbell, C. J. *et al.* Single-ion nuclear clock for metrology at the 19th decimal place. *Phys. Rev. Lett.* **108,** 120802 (2012).
14. Ruchowska, E. *et al.* Nuclear structure of $^{229}$Th. *Phys. Rev. C* **73,** 044326 (2006).
15. Karpeshin, F. F. & Trzhaskovskaya, M. B. Impact of the electron environment on the lifetime of the $^{229}$Th$^m$ low-lying isomer. *Phys. Rev. C* **76,** 054313 (2007).
16. Tkalya, E. V., Schneider, C., Jeet, J. & Hudson, E. R. Radiative lifetime and energy of the low-energy isomeric level in $^{229}$Th. *Phys. Rev. C* **92,** 054324 (2015).
17. Peik, E., Zimmermann, K., Okhapkin, M. & Tamm, C. Prospects for a nuclear optical frequency standard based on thorium-229. In *Proc. 7th Symp. on Frequency Standards and Metrology* (ed. Maleki, L.) 532–538 (World Scientific, 2009).
18. Zimmermann, K. *Experiments Towards Optical Nuclear Spectroscopy with Thorium-229*. PhD thesis, Univ. Hannover (2010).
19. Campbell, C. J., Radnaev, A. G. & Kuzmich, A. Wigner crystals of $^{229}$Th for optical excitation of the nuclear isomer. *Phys. Rev. Lett.* **106,** 223001 (2011).
20. Tkalya, E. V., Zherikhin, A. N. & Zhudov, V. I. Decay of the low-energy nuclear isomer $^{229}$Th$^m$(3/2$^+$, 3.5 $\pm$ 1.0 eV) in solids (dielectrics and metals): a new scheme of experimental research. *Phys. Rev. C* **61,** 064308 (2000).
21. Jeet, J. *et al.* Results of a direct search using synchrotron radiation for the low-energy $^{229}$Th nuclear isomeric transition. *Phys. Rev. Lett.* **114,** 253001 (2015).
22. Kazakov, G. A. *et al.* Performance of a $^{229}$thorium solid-state nuclear clock. *New J. Phys.* **14,** 083019 (2012).
23. Stellmer, S., Schreitl, M. & Schumm, T. Radioluminescence and photoluminescence of Th:$CaF_2$ crystals. *Sci. Rep.* **5,** 15580 (2015).
24. Nicholson, T. L. *et al.* Systematic evaluation of an atomic clock at $2 \times 10^{-18}$ total uncertainty. *Nature Commun.* **6,** 6896 (2015).
25. Flambaum, V. V. Enhanced effect of temporal variation of the fine structure constant and the strong interaction in $^{229}$Th. *Phys. Rev. Lett.* **97,** 092502 (2006).
26. Hayes, A. C., Friar, J. L. & Mller, P. Splitting sensitivity of the ground and 7.6 eV isomeric states of $^{229}$Th. *Phys. Rev. C* **78,** 024311 (2008).
27. Litvinova, E., Feldmeier, H., Dobaczewski, J. & Flambaum, V. Nuclear structure of lowest $^{229}$Th states and time-dependent fundamental constants. *Phys. Rev. C* **79,** 064303 (2009).
28. Rellergert, W. G. *et al.* Constraining the evolution of the fundamental constants with a solid-state optical frequency reference based on the $^{229}$Th nucleus. *Phys. Rev. Lett.* **104,** 200802 (2010).
29. Matinyan, S. Lasers as a bridge between atomic and nuclear physics. *Phys. Rep.* **298,** 199–249 (1998).
30. Raeder, S. *et al.* Resonance ionization spectroscopy of thorium isotopes — towards a laser spectroscopic identification of the low-lying 7.6 eV isomer of $^{229}$Th. *J. Phys. B* **44,** 165005 (2011).
31. Sonnenschein, V., Raeder, S., Hakimi, A., Moore, I. D. & Wendt, K. The search for the existence of $^{229m}$Th at IGISOL. *Eur. Phys. J. A* **48,** 52 (2012).
32. Kazakov, G. A. *et al.* Prospects for measuring the $^{229}$Th isomer energy using a metallic magnetic microcalorimeter. *Nucl. Instrum. Methods A* **735,** 229–239 (2014).
33. Irwin, G. M. & Kim, K. H. Observation of electromagnetic radiation from deexcitation of the $^{229}$Th isomer. *Phys. Rev. Lett.* **79,** 990–993 (1997).
34. Richardson, D. S., Benton, D. M., Evans, D. E., Griffith, J. A. R. & Tungate, G. Ultraviolet photon emission observed in the search for the decay of the $^{229}$Th isomer. *Phys. Rev. Lett.* **80,** 3206–3208 (1998).
35. Utter, S. B. *et al.* Reexamination of the optical gamma ray decay in $^{229}$Th. *Phys. Rev. Lett.* **82,** 505–508 (1999).
36. Zhao, X. *et al.* Observation of the deexcitation of the $^{229m}$Th nuclear isomer. *Phys. Rev. Lett.* **109,** 160801 (2012).
37. Peik, E. & Zimmermann, K. Comment on "Observation of the deexcitation of the $^{229m}$Th nuclear isomer". *Phys. Rev. Lett.* **111,** 018901 (2013).
38. Yamaguchi, A. *et al.* Experimental search for the low-energy nuclear transition in $^{229}$Th with undulator radiation. *New J. Phys.* **17,** 053053 (2015).
39. Peik, E. & Okhapkin, M. Nuclear clocks based on resonant excitation of $\gamma$-transitions. *C. R. Phys.* **16,** 516–523 (2015).
40. Swanberg, E. *Searching for the Decay of $^{229m}$Th.* PhD thesis, Univ. California Berkeley (2012).
41. von der Wense, L., Thirolf, P. G., Kalb, D. & Laatiaoui, M. Towards a direct transition energy measurement of the lowest nuclear excitation in $^{229}$Th. *J. Instrum.* **8,** P03005 (2013).
42. Barci, V., Ardisson, G., Barci-Funel, G., Weiss, B. & El Samad, O. Nuclear structure of $^{229}$Th from $\gamma$-ray spectroscopy study of $^{233}$U $\alpha$-particle decay. *Phys. Rev. C* **68,** 034329 (2003).
43. Neumayr, J. B. *et al.* Performance of the MLL-ion catcher. *Rev. Sci. Instrum.* **77,** 065109 (2006).
44. Jongma, R. T., Rasing, T. & Meijer, G. Two-dimensional imaging of metastable CO molecules. *J. Chem. Phys.* **102,** 1925–1933 (1995).
45. von der Wense, L., Seiferle, B., Laatiaoui, M. & Thirolf, P. G. Determination of the extraction efficiency for $^{233}$U source $\alpha$-recoil ions from the MLL buffer-gas stopping cell. *Eur. Phys. J. A* **51,** 29 (2015).
46. Goruganthu, R. R. & Wilson, W. G. Relative electron detection efficiency of microchannel plates from 0–3 keV. *Rev. Sci. Instrum.* **55,** 2030–2033 (1984).
47. Sakharov, S. L. On the energy of the 3.5 eV level in $^{229}$Th. *Phys. At. Nucl.* **73,** 1–8 (2010).
48. Schwarz, M. *et al.* Cryogenic linear Paul trap for cold highly charged ion experiments. *Rev. Sci. Instrum.* **83,** 083115 (2012).
49. Mårtensson, N. *et al.* A very high resolution electron spectrometer. *J. Electron Spectrosc.* **70,** 117–128 (1994).
50. Liao, W. T., Pálffy, A. & Keitel, C. H. Nuclear coherent population transfer with X-ray laser pulses. *Phys. Lett. B* **705,** 134–138 (2011).


**Acknowledgements** We acknowledge discussions with D. Habs, T. W. Hänsch, T. Udem, T. Lamour, J. Weitenberg, A. Ozawa, E. Peters, J. Schreiber, P. Hilz, T. Schumm, S. Stellmer, F. Allegretti, P. Feulner, J. Crespo, M. Schwarz, L. Schmöger, P. Micke, C. Weber, P. Bolton and K. Parodi. We thank T. Lauer for the Ti sputtering of the Si wafers and the MPQ for the temporary loan of the MCP detector. We thank I. Cortrie, L. Black and J. Soll for graphic design support. This work was supported by DFG (Th956/3-1), by the European Union's Horizon 2020 research and innovation programme under grant agreement 664732 "nuClock" and by the LMU department of Medical Physics via the Maier-Leibnitz Laboratory.

**Author Contributions** L.v.d.W., B.S. and P.G.T. performed the experiments. M.L. and J.B.N. did preparatory experimental work. H.-J.M. and H.-F.W. produced the radioactive source 1. C.M., J.R., K.E., C.E.D., N.G.T. and L.v.d.W. produced the radioactive sources 2 and 3. L.v.d.W., P.G.T. and B.S. wrote the manuscript with input from all authors.

**Author Information** Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to L.v.d.W. (L.Wense@physik.uni-muenchen.de).

## METHODS

**$^{233}U$ and $^{234}U$ $\alpha$-recoil ion sources.** Three different sources were employed in these experiments. Source 1 consists of about 200 kBq $^{233}U$ ($UF_4$), evaporated in vacuum from a tantalum heater lined with a vitreous carbon crucible[51] onto a 20-mm-diameter stainless-steel plate. The preparation was performed in the former hot-lab facility of the LMU Munich[52]. The $UF_4$ layer thickness is $360 \pm 20$ nm, leading to a recoil efficiency of about 5.3% for $^{229}Th$. The source material was not chemically purified before evaporation. As the material was produced around 1969, a significant ingrowth of short-lived daughter nuclides has occurred since then. An unavoidable fraction of $^{232}U$ contamination was determined by $\gamma$ spectroscopy to be $(6.1 \pm 0.3) \times 10^{-7}$ at the time of material production[45].

Source 2 consists of $270 \pm 10$ kBq $^{234}U$, deposited by molecular plating[53] onto the surface of a Ti-sputtered Si wafer of 100 mm diameter. It has a thickness of 0.5 mm with a 100 nm thick layer of sputtered titanium. The active surface area of the source is 90 mm in diameter, leaving a 12 mm diameter unplated region in the centre.

Source 3 is a newly available $^{233}U$ source of about 290 kBq. Just like source 2, it was deposited by molecular plating with 90 mm diameter onto the surface of a Ti-sputtered Si wafer of 100 mm diameter. Because of the smaller source thickness, the thorium extraction rate was increased by a factor of about 13.5 compared to source 1. The source 3 material was chemically purified before deposition by ion-exchange chromatography to remove the $^{233}U$ and $^{232}U$ daughter nuclides. A relative purification factor of $\geq$250 was found, based on a comparison of $\gamma$-energy spectra of the source material before and after chemical purification.

**Buffer-gas stopping cell.** The uranium source is mounted in the buffer-gas stopping cell[43] (Extended Data Fig. 2) and acts as an electrode of the ion-extraction system (39 V offset voltage). The $\alpha$-recoil ions, which possess a kinetic energy of up to 84.3 keV for $^{229}Th$, are stopped in 40 mbar of ultra-pure helium. In order to guarantee the required cleanliness of the buffer gas, helium with a purity of 99.9999% is used, which is further purified by catalytic purification (SAES Getters, MonoTorr, phase 2) and a cryotrap filled with liquid nitrogen. The gas tubing was electropolished and the cell chamber was built to UHV standards, bakeable up to 180 °C. A typical background pressure of $P \leq 3 \times 10^{-10}$ mbar is achieved. This high cleanliness allows for the extraction of $^{229}Th$ even in the 3+ charge state[45].

The buffer-gas stopping cell also houses the RF+ DC funnel system, consisting of 50 ring electrodes of 0.5 to 1 mm thickness, converging from 115 mm to 5 mm inner diameter. RF- and DC voltages are applied to this electrode structure. The applied RF voltages are 220 $V_{pp}$ at 850 kHz, varying in phase by 180° between neighbouring electrodes. This leads to a repelling force, preventing the recoil ions from charge exchange at the electrodes. In parallel, a DC voltage gradient of 4 V cm$^{-1}$ is applied by a voltage-divider chain (35 V to 3 V), guiding the ions through the buffer-gas background towards the buffer-gas stopping-cell exit. The latter consists of a supersonic Laval nozzle (2 V offset) with a 0.6 mm diameter nozzle throat. In this way, supersonic velocities of the helium gas flow are achieved and the $\alpha$-recoil ions are extracted from the buffer-gas stopping cell together with the helium carrier gas.

**RFQ ion guide and cooler.** Following the buffer-gas stopping cell, the ions are injected into an RFQ system, which consists of four rods with 11 mm diameter, with a 10 mm distance between opposite rods. For ion guiding, an RF field of 200 $V_{pp}$ at 880 kHz is applied. Each rod is divided into 12 segments and the overall length of the system is 33 cm. Because of the segmentation we can apply an individual DC voltage to each segment, thereby establishing a voltage gradient of 0.1 V cm$^{-1}$ (1.8 V to 0 V) to drag ions through the remaining helium buffer-gas background of about $10^{-2}$ mbar, or to store the ions in the RFQ. This background pressure is used for phase-space cooling of the recoil ions, which leads to a sub-millimetre diameter recoil-ion beam at the RFQ exit. By voltage control of the last RFQ electrode, the ion beam can optionally be pulsed.

**Quadrupole mass-separator.** Following the RFQ, the $\alpha$-recoil ions are mass separated in a quadrupole mass-separator (QMS)[54]. The QMS consists of four rods with 18 mm rod diameter and 15.96 mm inner rod distance. The length is 30 cm, with an additional 5 cm at the entrance and exit acting as Brubaker lenses[55]. At the resonance frequency of 925 kHz, an RF amplitude of 600.5 $V_{pp}$ and a DC voltage of 50.15 V is required for the extraction of $^{229}Th^{3+}$ (901.5 $V_{pp}$ and 75.23 V for the 2+ charge state, respectively). A voltage offset of $-2$ V is applied to the whole system. With this device, a transmission efficiency exceeding 70% with a mass resolving power of $m/\Delta m = 150$ can be achieved.

Prior to any isomer detection, the QMS is calibrated in order to extract ions of wanted mass-to-charge ratio. The mass spectrum (Fig. 4a) is well known from earlier measurements[45], where the correctness of the peak assignment was proven by parallel detection with a silicon detector for $\alpha$ spectroscopy and an MCP detector. Given this mass spectrum, the QMS is calibrated by performing ion-impact profile measurements (Extended Data Fig. 3 lower panel) with the beam-imaging MCP detector (Beam Imaging Solutions, BOS-75-FO), when operating the detector at a surface voltage of about $-900$ V. Consequently the impact of the transmitted ions is detectable (due to their kinetic energy of 1.8 to 2.7 keV, depending on the charge state). During calibration, care has to be taken not to contaminate the detector surface with short-lived daughter nuclides. For this purpose, the scans are always started at higher masses (above $^{233}U$) and stopped when the $^{229}Th^{2+}$ mass peak is reached.

**Triodic extraction system.** Behind the QMS, the ions are guided by a triodic electrode structure consisting of three ring electrodes in a nozzle-like shape. The first electrode acts as an aperture electrode to shield the RF voltages of the QMS ($-2$ V). A voltage of $-62$ V is applied to the second electrode in order to extract the ions from the QMS. The third electrode with a 2-mm diameter opening shields the extraction voltage from the surroundings when applying $-22$ V. As a result ions are guided to the MCP detection system. A combined extraction and purification efficiency for $Th^{3+}$ of $(10 \pm 2)$% was determined behind the triodic extraction system[45]. Together with the 5.3% recoil efficiency of source 1, $(1.0 \pm 0.1) \times 10^3$ $^{229}Th^{3+}$ ions per second are extracted. A $(5.5 \pm 1.1)$% extraction efficiency was obtained for $Th^{2+}$, resulting in $(5.8 \pm 0.6) \times 10^2$ extracted $Th^{2+}$ ions per second. The total time for extraction is a few ms (3 to 5 ms were obtained as extraction times behind the RFQ[56]). Faster decays of nuclear excitations already take place in the buffer-gas stopping cell.

**MCP detection system.** The ions are collected directly on the surface of a micro-channel plate (MCP) detector[57] placed at 5 mm distance to the last electrode of the triodic extraction system (Fig. 2). The MCP detector (Beam Imaging Solutions, BOS-75-FO) consists of two MCP plates (chevron geometry, 25 μm channel diameter) with 75 mm diameter. The front surface is CsI-coated. The two plates are positioned in front of a vacuum-flange-mounted optic fibre-glass window, which is coated with a phosphor layer. During extraction, the MCP is operated at an He pressure of $10^{-6}$ mbar, and typical voltages of $-25$ V and $+1{,}900$ V are applied to the front and the back sides of the MCP, respectively. A voltage of $+6{,}000$ V is applied to the phosphor screen, which is monitored through the optic fibre-glass window by a CCD camera (FL2-14S3M-C, PointGrey) with a zoom lens (Computar M2514MP2, 25 mm, C-mount). The distance between the window and the CCD camera is about 30 cm, leading to a field of view of 100 mm by 75 mm. The outer region of the optical window is covered by a 20-mm diameter aperture in order to cover arcing effects from the detector's side. The camera is mounted onto an optical rail, which is placed in a light-tight housing.

Owing to the expected short isomeric lifetime in neutral thorium, it is important to allow for $^{229m}Th$ decay detection during ion accumulation, which affords probing even for decays that would occur simultaneously with charge exchange on the MCP surface. For this purpose, the MCP is operated with a surface voltage of $-25$ V. In this way, the thorium ions are collected at low kinetic energy (50–75 eV, depending on the charge state) in a soft landing onto the MCP surface. The remaining kinetic energy of the ions as well as the energy carried by the ions in the form of the charge state does not lead to a significant signal on the MCP surface[58]. Most of the energy in these processes is transferred to phonons at the point of impact with the surface[59]. No ion-impact signal was detected with an MCP surface voltage above $-40$ V (that is, a negative voltage with magnitude below 40 V).

Relatively little is known about the detection efficiency of MCPs for low-energy electrons (the ionization potential of thorium is 6.31 eV, thus an IC electron kinetic energy of about 1.5 eV remains, given a 7.8-eV isomeric transition). Applying the model discussed in ref. 46, a decrease in detection efficiency to 2.9% of the maximum value (at about 300 eV kinetic energy) is predicted for incident electrons of 1.5 eV energy. Assuming a maximum detection efficiency of 50% (corresponding to the channel open area of the MCP), an absolute detection efficiency of about 1.5% is expected. 2% of the 1,000 $^{229}Th^{3+}$ ions which are extracted per second are predicted to be in the isomeric state[42]. Comparing this with the detected isomeric-decay count rate of 0.25 per second leads to an experimentally obtained detection efficiency of 1.3%, which is in good agreement with our expectation.

**Image readout.** For readout of the MCP signal, the CCD chip (Sony ICX267 CCD, $4.65 \times 4.65\,\mu m^2$ pixel size, $1{,}384 \times 1{,}032$ pixels) was exposed for 4 s for each frame. In these frames, single events of the MCP detector can clearly be distinguished from the CCD intrinsic background (noise and hot pixels) by size and intensity. A Matlab program is applied to determine the position of each individual event. These events are then added for a chosen number of frames (typically 500 for 2,000 s integration time) to obtain one single image. Appropriate choice of the filter parameters of the program is tested by an individual control of 50 images. The loss of events due to low signal intensity on the phosphor screen or due to spatial overlap is found to be negligible. Only a minor amount of CCD intrinsic noise is not adequately filtered. By applying this type of image readout, the background is dominated by the MCP intrinsic dark-count rate of about 0.01 counts s$^{-1}$ mm$^{-2}$.

**Code availability.** All programs used for image read out are available by email on request without restriction. Requests for program codes should be addressed to L.v.d.W. (L.Wense@physik.uni-muenchen.de).

**Signal comparison.** CCD camera images of the phosphor screen reveal features that enable us to distinguish type or origin of signals. Signals of different origin are shown in Extended Data Fig. 4. Each image corresponds to 4 s exposure time of the CCD chip (that is, one frame). A wanted ion species was chosen by mass-to-charge selection with the QMS. Extended Data Fig. 4a shows $\alpha$ decays on the MCP surface occurring within 5 min after extraction of $^{221}Fr^{2+}$ ($t_{1/2} = 286$ s). Very large and intense signals are seen, with an average diameter of about 1 mm. Extended Data Fig. 4b shows $\beta$ decays occurring within 45 min after extraction of $^{209}Pb^{2+}$ ($t_{1/2} = 3.25$ h). The signals are significantly smaller and less intense than those caused by $\alpha$ decays. The typical signal diameter is about 0.6 mm. Extended Data Fig. 4c shows signals caused by the isomeric decay of $^{229}$Th starting to occur simultaneously with the accumulation of $^{229}Th^{3+}$ on the MCP surface. The signals appear small and of low intensity with a typical signal diameter of about 0.3 mm. They are slightly smaller than the signals caused by $\beta$ decays, and clearly distinguishable from the $\alpha$ events. Finally, signals caused by the isomeric 76 eV IC decay of $^{235m}$U ($t_{1/2} = 26$ min) are shown in Extended Data Fig. 4d, taken within 30 min of extraction of $^{235}U^{2+}$. They are comparable with the isomeric-decay signals of $^{229}$Th.

**Half-life measurements.** Two different half-life measurements are implemented. The first measurement leads to an upper limit for the isomeric half-life in neutral thorium. To obtain this limiting value, a pulsed $^{229(m)}Th^{2+}$ ion beam is produced by applying a gate voltage of 0.5 V to the last RFQ electrode. The gate is opened for 500 ms and is then closed for 1,700 ms, while ions are accumulated in the RFQ continuously (a maximum storage time of about 1 min is obtained for $Th^{2+}$). Strong ion pulses are produced when the QMS is set to extract $^{229}Th^{2+}$. This is controlled by applying an MCP surface voltage of about −900 V, yielding strong ion-impact signals. The CCD camera acquires images of 1 s exposure time only when the beam gate is closed. To ensure that the gate is actually closed, the camera is started 500 ms after applying the gate voltage. The camera is stopped after 1,200 ms in parallel to the gate opening, in order to acquire one image per pulse. It is reconfirmed that the camera does not acquire pictures at times of ion impact. By this sequence 1,200 frames (corresponding to 1,200 s total exposure time) are evaluated. No signal is obtained, which means that the isomer half-life must be below 1 s, allowing for charge exchange of the $^{229}Th^{2+}$ ions on the MCP surface.

In a second measurement, a lower limit of the isomeric lifetime in $^{229}Th^{2+}$ is found. For this purpose, $^{229(m)}Th^{2+}$ ions are stored in the RFQ, by applying a gating voltage of 5 V to the last RFQ electrode. After storage, the ion cloud is accelerated onto the MCP surface to examine survival of the isomeric state by detected internal conversion. The half-lives that can be probed by this method are limited by the storage times of $Th^{2+}$ in the RFQ. A one minute storage time is easily accessible without significant ion loss. For this measurement the ions are accumulated for 10 s in the RFQ, where they are stored. After 10 s the $^{233}$U source offset is reduced to 0 V, preventing additional recoil ions from leaving the buffer-gas stopping cell. Then the ions are stored for one minute in the RFQ, waiting for the isomeric decay to occur. Afterwards, the gate voltage is also reduced to 0 V and the isomeric decay is read from the MCP detector. To reduce the dark count, the CCD camera is triggered to only acquire images when the ions are released. In this way, 200 pulses are evaluated with 3 imaged frames per pulse (4 s exposure time for each frame). A clear signal is seen when the QMS is set to extract $^{229}Th^{2+}$, from which is inferred a half-life greater one minute. To eliminate signal contribution from a long-lived $\beta^-$ emitter, which might have populated the $^{229}Th^{2+}$ mass peak by molecular formation (for example, oxides), a measurement of the background rate is performed afterwards for 1 h and no signal is obtained.

**Sample size.** No statistical methods were used to predetermine sample size.

**Exclusion of nuclear background from $^{229}Th^{2+}$ and $^{229}Th^{3+}$ signal comparison.** All potential background contributions together with the relevant means of exclusion (nos 1–7) are listed in Extended Data Table 1. The present section relates to Extended Data Table 1 no. 4.

It has been discussed that the parallel occurrence of the signal in the 2+ and 3+ charge states (Fig. 4b, first row) is already sufficient to exclude nuclear background as potentially originating from short-lived daughter nuclides. The reason is that only thorium can be expected to be extracted to a significant extent in the 3+ charge state, due to its low third ionization potential of only 18.3 eV (see Extended Data Table 2), which is below the first ionization potential of He (24.6 eV). Thus, during stopping in the helium environment, it is energetically favourable for the electrons to stay attached to the helium atoms instead of reducing the thorium 3+ charge state. Experimentally, a reduced extraction for all short-lived daughter nuclides (of atomic number $Z = 88$ or below) by three to four orders of magnitude is found in the 3+ compared to the 2+ charge state[45]. In the case of signals not caused by $^{229}$Th, the same reduction of signal intensity would be expected when comparing the 2+ and 3+ charge states. This, however, is not observed. For completeness, all ionization potentials[60] for the elements which are potentially contained in the source material are listed in Extended Data Table 2. The heavier elements also possess low ionization potentials, but they cannot explain the observed signal because their half-lives are significantly longer. Further, mass-peak shifts from heavier to lighter masses cannot be explained by the population of molecular side-bands.

**Signal comparison from chemically purified and unpurified $^{233}$U sources.** This section relates to Extended Data Table 1 no. 5. Isomeric-decay measurements were also performed with the new chemically purified $^{233}$U source of 90 mm diameter (source 3, 290 kBq). The thorium ions were collected in the 2+ and 3+ charge states for 2,000 s, while detection was performed in parallel. Compared to source 1, these measurements resulted in ∼13.5 times higher isomeric count rate (∼3.4 counts $s^{-1}$). This enhancement occurs because of the reduced source thickness, leading to a higher $\alpha$-recoil efficiency. The results of these measurements are shown in Fig. 4b, third row.

If signals were caused by a decay of any of the short-lived daughter nuclides, a considerable decrease in signal intensity compared to source 1 should occur, due to the chemical purification factor of more than 250. The fact that this is not the case further serves to exclude radioactive decay of short-lived isotopes as a signal contribution.

**Signal appearance and the $^{229m}$Th half-life limit.** This section relates to Extended Data Table 1 no. 6. It was shown that the uniquely strong signal shape excludes any $\alpha$ decay as a contributor to the observed decay events (see Extended Data Fig. 4). This information, in combination with the observed short decay half-life (in the subsecond region), is already sufficient to exclude any nuclear origin of the signal except for the isomeric decay of $^{229}$Th.

While the uranium-source material predominantly consists of $^{233}$U, in source 1, trace amounts of other nuclides are also included ($^{232}$U, $^{238}$Pu, $^{239}$Pu, $^{231}$Pa) together with their decay daughters[45]. Even further nuclides could potentially be present, although they have not been experimentally observed. A complete list of nuclides potentially contained in the source material (produced by neutron irradiation in a nuclear reactor) is shown in Extended Data Fig. 5. Their half-lives and decay branching ratios are also listed[1]. For completeness, all populated nuclides are shown, even if their activity can be assumed to play only a negligible role due to their small branching ratio or due to a long half-life of the mother nuclide. A complete list of potentially contributing isomers is given in Extended Data Table 3, together with their corresponding excitation energies and half-lives[1]. Note that excited states with half-lives in the microsecond range or shorter do not have to be considered, because the extraction time from the source is in the millisecond range[56].

As can be inferred, there is no pure $\beta$ emitter or isomer contained in this list that could potentially explain the detected signal, except for the 0.8 s isomeric state in $^{207}$Pb. This isomeric state is, however, populated only by a fraction of $8.1 \times 10^{-6}$ from the $\alpha$ decay of $^{211}$Po, which itself is not part of the main decay chains.

Furthermore, the $^{229}$Th isomeric transition is the only known nuclear transition which is expected to reveal the observed strong dependence of its half-life on the electronic environment. Thus the detected signal cannot be explained by any nuclear decay other than the decay of the $^{229}$Th first excited nuclear state.

**Search for $\alpha$ and $\beta$ decays using Si and $LN_2$ cooled Si(Li) detectors.** This section relates to Extended Data Table 1 no. 7. To further substantiate the evidence, the extracted ions (when operating the QMS for collection of $^{229}Th^{2+}$ or $^{229}Th^{3+}$) are directly accumulated on the surface of two different silicon detectors. The first detector is optimized for $\alpha$-particle detection in order to provide a further exclusion of $\alpha$ decays as a cause of the detected signals. The second detector is used in order to exclude $\beta$ decays or high-energy internal conversion electrons.

For the exclusion of $\alpha$ decays, an ion-implanted silicon charged particle detector (Ametek, BU-014-150-100) is used. This detector is mounted directly behind the extraction triode at about 5 mm distance, replacing the CsI-coated MCP detector. A charge sensitive preamplifier (CSTA) and a shaping amplifier (Ortec, model 571) are used for signal processing. The spectra are acquired by a multi-channel analyser (Amptek, MCA-8000 A). The detector is operated at a 20 V bias voltage and a −10 V offset is applied to the whole system in order to collect the ions directly on the detector surface. To allow also for mass scans as required for the calibration of the QMS, an MCP detector (Hamamatsu, type F2223) is mounted sideways at 90° to the extraction triode. During QMS calibration, a surface voltage of −2,000 V is applied to the MCP, which is sufficiently high to attract the ions in spite of its off-axis position. After the QMS has been set to extract the desired ion species, the MCP surface voltage is reduced to 0 V, so ions are collected on the Si-detector surface. In this way, four different measurements were performed, each with 2 h acquisition time: one during the extraction of $^{213}Bi^{2+}$ (2.0 counts $s^{-1}$) in order to prove the functionality of the detector system, one dark count measurement ($5.7 \times 10^{-3}$ counts $s^{-1}$), one during the extraction of $^{229}Th^{2+}$ ($6.0 \times 10^{-3}$ counts $s^{-1}$) and one during the extraction of $^{229}Th^{3+}$ ($5.3 \times 10^{-3}$ counts $s^{-1}$).

The corresponding spectra are shown in Extended Data Fig. 6a–d. As expected, no entries above background are visible in the energy range where $\alpha$-particles from $^{229}$Th would appear (4.7–5.1 MeV) during the extraction of $^{229}$Th in the 2+ or 3+ charge state, as the half-life of the $^{229}$Th $\alpha$ decay is 7,932 years and thus practically no decays occur within the duration of these comparatively short measurements. The fact that no line from any $\alpha$-decaying nucleus is visible in the spectra allows exclusion of $\alpha$-decays as the origin of the 0.25 counts $s^{-1}$ signal measured on the MCP in the search for the isomeric decay of $^{229m}$Th: had this signal originated from $\alpha$-decays, a total of about 1,800 counts should have be seen in a 2 h measurement with the Si detector, which would have been easily visible.

While $\beta$ decays as a signal contribution have already been excluded by half-life arguments, a further way to exclude them is given by direct $\beta$ detection. For this purpose a liquid nitrogen cryogenically cooled Si(Li) detector (Canberra, type ESLB-3000-300) is used, replacing the above mentioned Si detector. It is operated in combination with a preamplifier with a cooled FET stage (Eurisys Measures, PSC 761) and a shaping amplifier (Ortec, model 572). Again spectra are acquired by a multi-channel analyser (Amptek, MCA-8000 A). A bias voltage of −400 V is applied to the front surface, such that no further offset is required. The detector is mounted at 5 mm distance from the triodic extraction system. Four different measurements were performed, each with 10 h acquisition time: one dark-count measurement (0.47 counts $s^{-1}$), one during the extraction of $^{229}Th^{2+}$ (0.44 counts $s^{-1}$), one during the extraction of $^{229}Th^{3+}$ (0.48 counts $s^{-1}$) and one during the extraction of $^{209}Pb^{2+}$ (2.13 counts $s^{-1}$), the last to prove the functionality of the detection system. If the detected signals were $\beta$ decays or high-energy internal-conversion electrons, the expected enhancement of the integrated signal rate of $0.25 \pm 0.1$ counts $s^{-1}$ (for source 1) would have been detected easily.

**Prospects for energy determination.** A precise determination of the isomer's energy is one of the most important prerequisites for the development of a nuclear frequency standard. The direct detection of the isomeric decay opens new perspectives for such an energy determination. In the presented work, the IC decay channel in the neutral thorium atom is investigated. Any energy determination based on this direct detection will require energy spectroscopy of the IC electrons emitted in the isomeric decay.

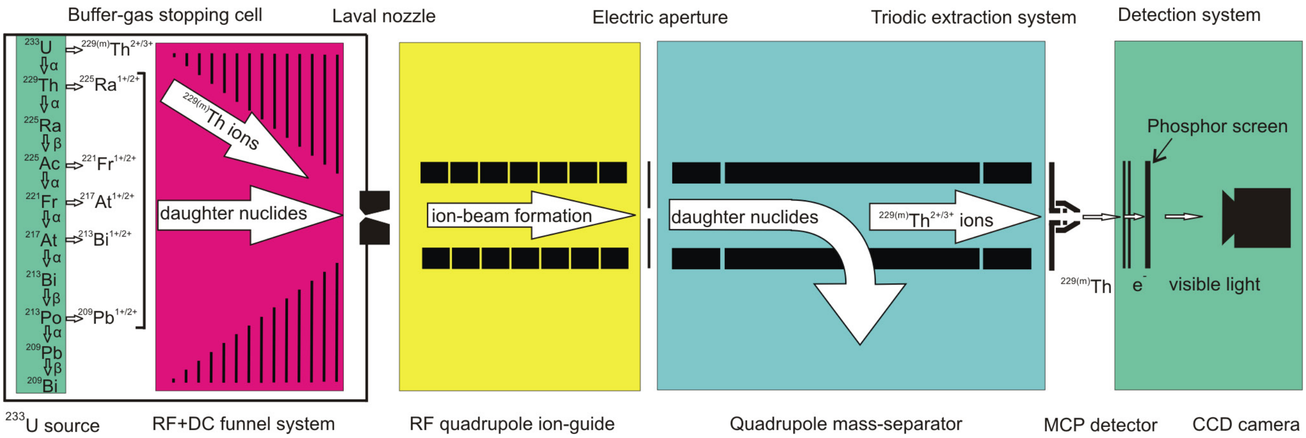
Several techniques for electron-energy spectroscopy of different precisions and complexities are known. The highest known precision is provided by hemispherical electron energy analysers, which possess resolutions in the range of a few meV (ref. 49). While being also among the most complex devices for spectroscopy, there is a trade-off between energy resolution and signal contrast. This problem can be solved by ion-beam pulsing. When applying an RFQ buncher, ion bunches with a pulse length of a few tens of nanoseconds can be produced[61]. These bunches are significantly shorter than the expected isomeric lifetime in the neutral thorium atom, which is predicted to be in the microsecond range. Such ion beam pulsing would not only allow the determination of the isomer's half-life in the neutral thorium atom, but also the suppression of any background by several orders of magnitude (depending on the exact isomeric half-life) if the electron detector is triggered in accordance with these pulses. This improvement in signal-to-background ratio will make high resolution electron spectroscopy applicable to the problem of energy determination of the isomeric state.

A considerably simpler sort of electron spectrometer is provided by retarding field analysers, which consist of a set of concentric hemispherical grids[62]. While this technique is significantly easier to apply, the achieved energy resolution is typically in the range of a few 100 meV. The expected low signal-to-background ratio of this technique will again make short ion-beam pulsing an important tool.

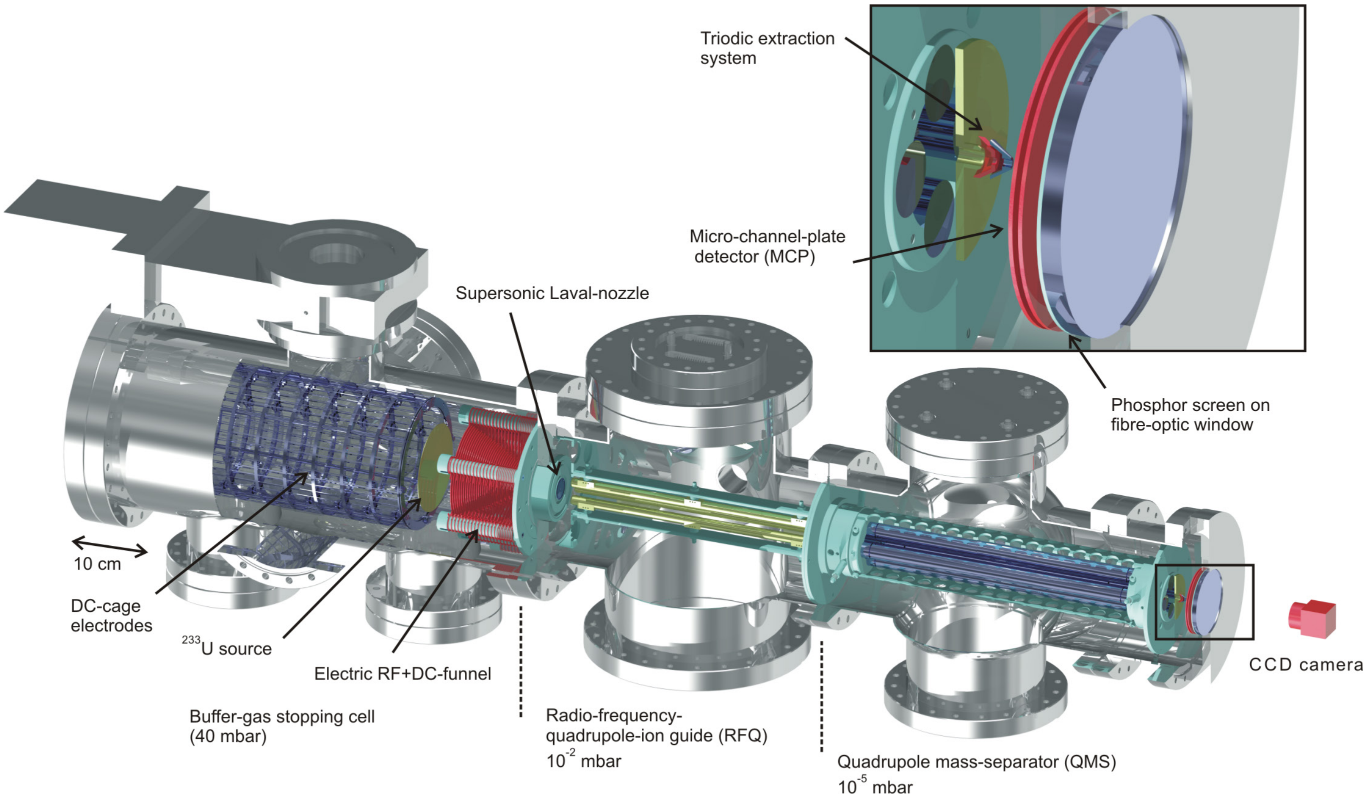
Independent of the applied technique for electron spectroscopy, charge exchange is required for the thorium ions in order to trigger the IC decay. In the simplest approach, this charge exchange is achieved by deposition of the thorium ions on a surface. This technique, however, is expected to influence the energy of the IC electrons as the work function of the surface material has to be considered[63]. In case of CsI, the coating material of the MCP used for all presented detections, the work function is 6.2 eV and thus close to the first ionization potential of thorium[64]. For this reason no drastic influence on the reported energy of the IC electrons is expected. For a precise energy determination, however, a careful investigation of surface influences is required. This must include the collection of the thorium ions on different surface materials with different work functions.

An alternative to the collection on a surface could be provided by collision with an atom beam. By crossing the thorium ion beam with a beam of, for example, caesium atoms, charge exchange will trigger the isomeric decay and could lead to an improved energy determination as no surface influences have to be considered[65].

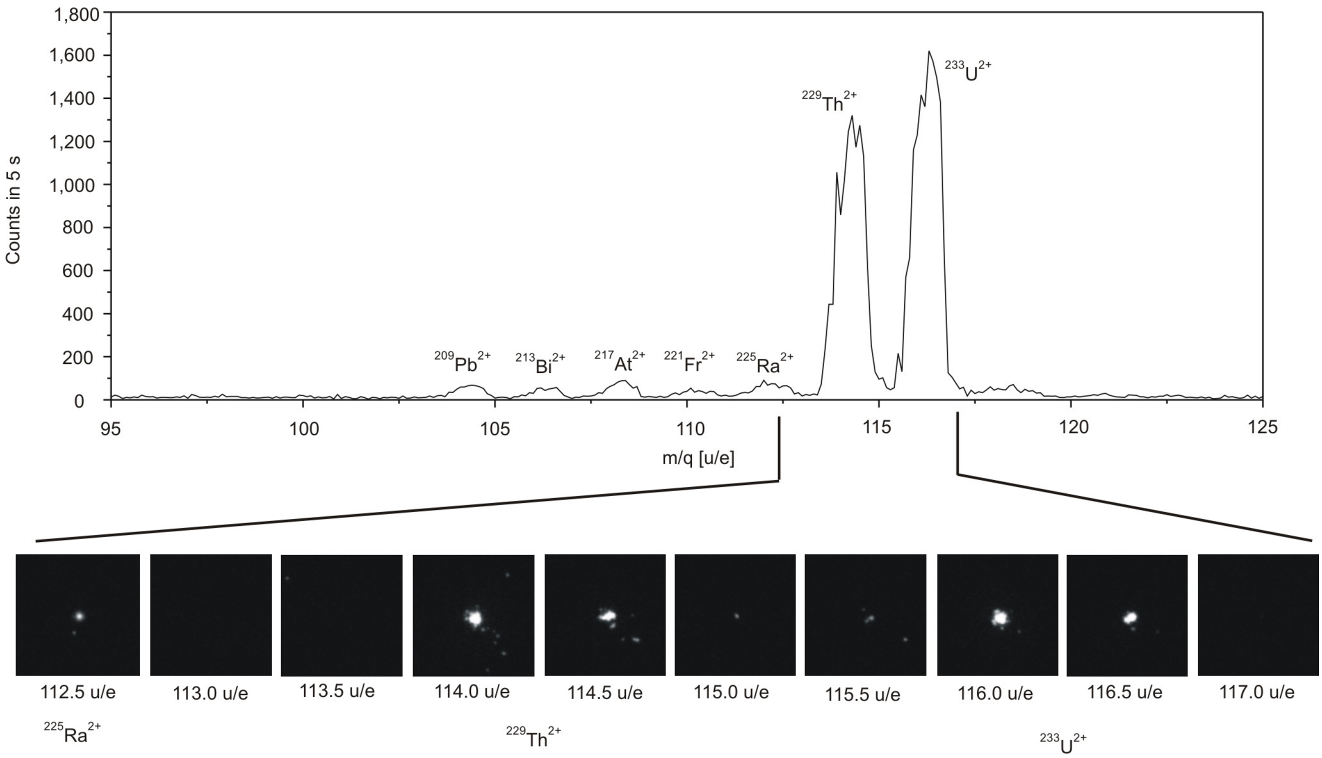

51. Maier, H. J., Grossmann, R. & Friebel, H. U. Radioactive targets for nuclear accelerator experiments. *Nucl. Instrum. Methods B* **56-57,** 926–932 (1991).
52. Grossmann, R., Maier, H. J. & Friebel, H. U. The new hot-lab facility for radioactive target preparation at the University of Munich. *Nucl. Instrum. Methods A* **397,** 39–45 (1997).
53. Vascon, A. *et al.* Elucidation of constant current density molecular plating. *Nucl. Instrum. Methods A* **696,** 180–191 (2012).
54. Haettner, E. *A Novel Radio Frequency Quadrupole System for SHIPTRAP and New Mass Measurements of rp Nuclides*. PhD thesis, Univ. Giessen (2011).
55. Brubaker, W. M. An improved quadrupole mass analyser. *Adv. Mass Spectrom.* **4,** 293–299 (1968).
56. Neumayr, J. B. *The Buffer-gas Cell and the Extraction RFQ for SHIPTRAP*. PhD thesis, Univ. Munich (2004).
57. Wiza, J. L. Microchannel plate detectors. *Nucl. Instrum. Methods* **162,** 587–601 (1979).
58. Rispoli, R. *et al.* ELENA MCP detector: absolute efficiency measurement for low energy neutral atoms. http://dx.doi.org/10.1117/12.926185 *Proc. SPIE* **8450,** 84505L (2012).
59. Bay, H. L., Winters, H. F., Coufal, H. J. & Eckstein, W. Energy transfer to a copper surface by low energy noble gas ion bombardment. *Appl. Phys. A* **55,** 274–278 (1992).
60. Kramida, A., Ralchenko, Yu. & Reader, J. NIST atomic spectra database (version 5.2) http://physics.nist.gov/asd (accessed 16 June 2015).
61. Plaß, W. R. *et al.* Isobar separation by time-of-flight mass spectrometry for low-energy radioactive ion beam facilities. *Nucl. Instrum. Methods B* **266,** 4560–4564 (2008).
62. Palmberg, P. W. Optimization of Auger electron spectroscopy in LEED systems. *Appl. Phys. Lett.* **13,** 183–185 (1968).
63. Hotop, H. Detection of metastable atoms and molecules. *Exp. Methods Phys. Sci. B* **29,** 191–215 (1996).
64. Poole, R. T., Jenkin, J. G., Liesegang, J. & Leckey, R. C. G. Electronic band structure of the alkali halides. I. Experimental parameters. *Phys. Rev. B* **11,** 5179–5189 (1975).
65. Yamakita, Y. *et al.* A highly sensitive electron spectrometer for crossed-beam collisional ionization: a retarding-type magnetic bottle analyzer and its application to collision-energy resolved Penning ionization electron spectroscopy. *Rev. Sci. Instrum.* **71,** 3042–3049 (2000).



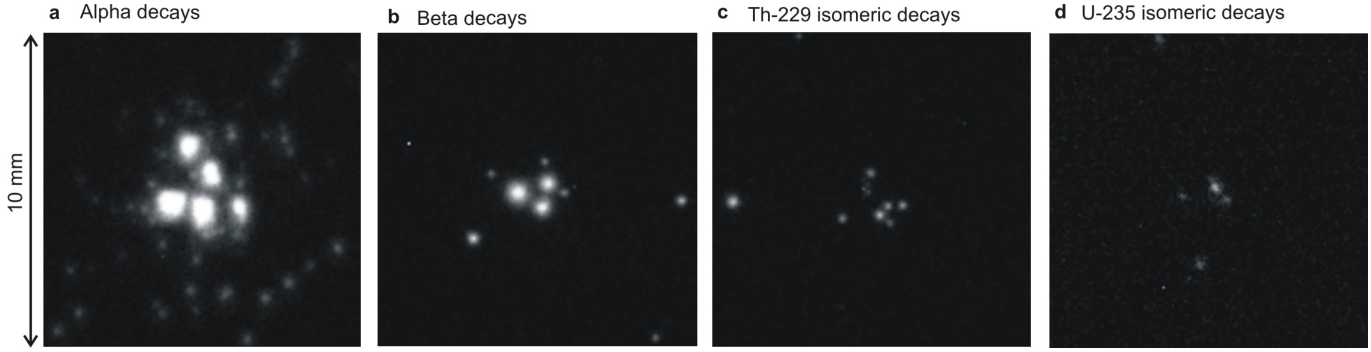

**Extended Data Figure 1 | Schematic of the experimental process.** Daughter nuclides of the $^{233}$U decay chain leave the $^{233}$U source owing to the kinetic recoil energy transferred to the nucleus during the $\alpha$ decay. Only those nuclides produced by $\alpha$ decay have enough kinetic recoil energy to leave the $^{233}$U source material efficiently. The maximum layer thickness through which recoiling nuclei can pass is a few tens of nanometres. The $\alpha$-recoil nuclei are thermalized with helium and extracted from the stopping cell. The process of electron capture during thermalization leads to the formation of ions in the 1+, 2+ or 3+ charge states. Subsequently, an ion beam is formed and purified with a quadrupole mass-separator such that only $^{229}$Th remains. The thorium ions are collected by soft landing on the surface of a microchannel-plate (MCP) detector and the isomeric decay is detected. Major components shown here are described in detail in Extended Data Fig. 2 legend.



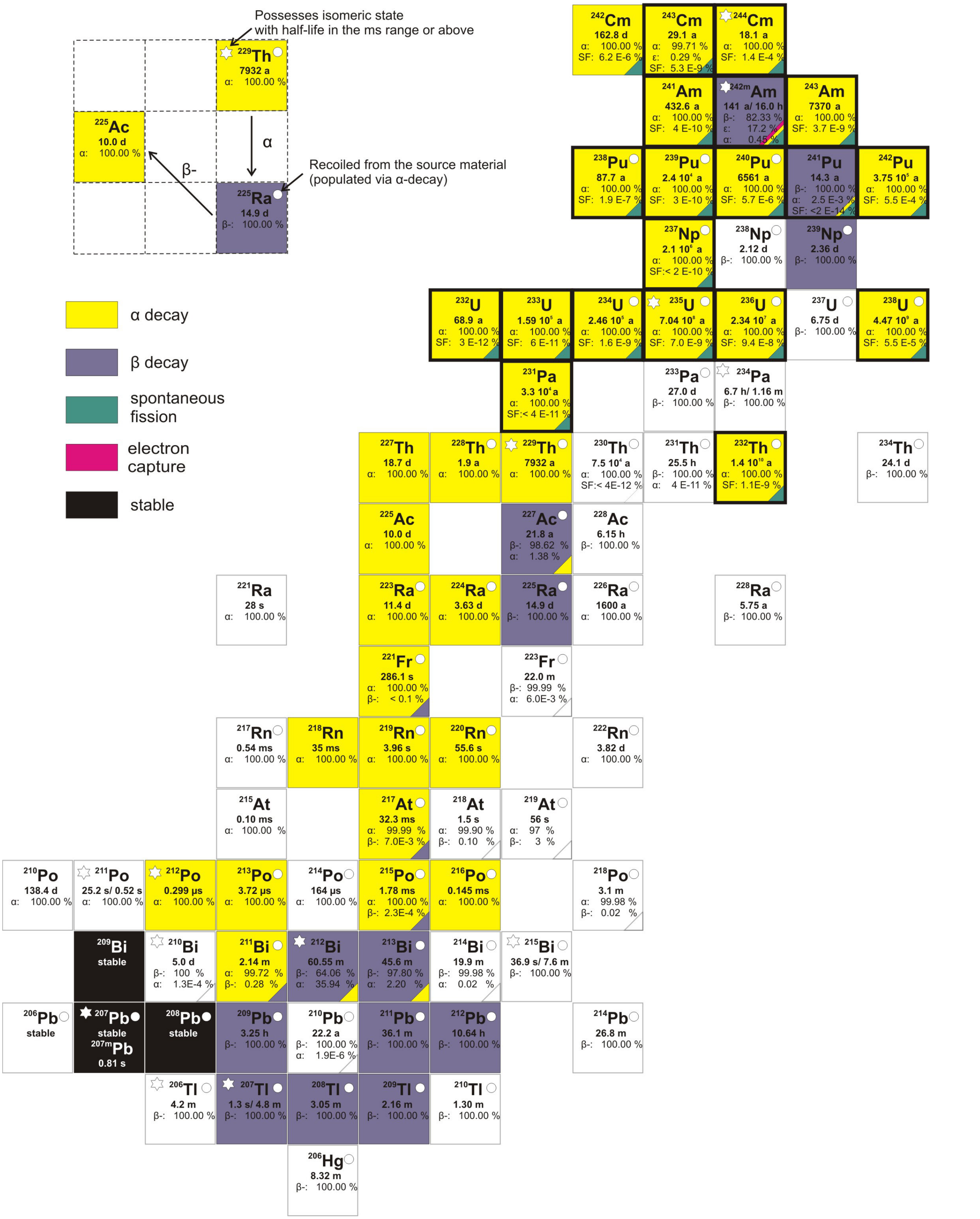

**Extended Data Figure 2 | Overview of the experimental setup.** The buffer-gas stopping cell houses the $^{233}$U source, which is mounted onto the front end of a DC cage electrode system[56]. The $^{229}$Th $\alpha$-recoil ions emitted from the source are stopped in the buffer-gas stopping cell filled with 40 mbar helium. These ions are then guided by an electric RF+ DC funnel system towards the exit of the stopping cell formed by a supersonic Laval nozzle, which injects them into a radio-frequency quadrupole (RFQ) ion guide, where an ion beam is formed by phase-space cooling due to the remaining helium pressure of $10^{-2}$ mbar. Following the RFQ, the ion beam is purified after a mass-to-charge separation with a quadrupole mass-separator (QMS). Behind the QMS a microchannel plate (MCP) allows for the detection of the low-energy internal conversion (IC) electrons emitted in the $^{229}$Th isomeric decay. Boxed area at right is shown magnified in inset.



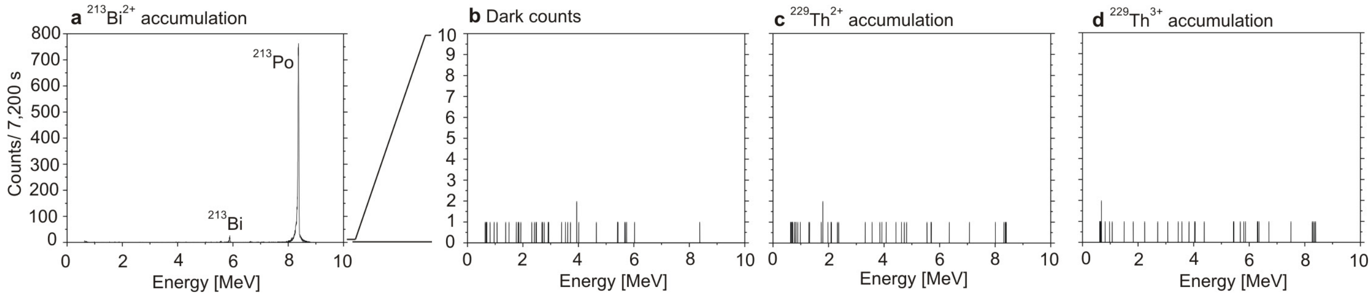

**Extended Data Figure 3 | Intensity profile measurements.** Upper panel, mass spectrum in the range of the 2+ ion species as performed with the chemically unpurified $^{233}$U source 1 and an MCP detector (Methods) operated in single-ion counting mode. Lower panel, ion impact profile measurement (−900 V MCP surface voltage, 1 s exposure time) performed with $^{233}$U source 1 and an MCP detector allowing for spatially resolved read-out (Methods). The $^{229}$Th and $^{233}$U mass peaks can clearly be separated.




**Extended Data Figure 4 | Different classes of decay events as observed during ion accumulation on the MCP surface.** In order to suppress any ion-impact signal, soft landing of the ions is guaranteed at −25 V MCP surface voltage. Single frames of 4 s exposure time are shown. The MCP detector used (Methods) allows for spatially resolved image read-out. The extracted ion species is chosen by mass-to-charge separation with the help of the QMS. **a**, Alpha decays originating from $^{221}$Fr. **b**, Beta decays originating from $^{209}$Pb. **c**, Isomeric decay of $^{229}$Th. **d**, Isomeric decay of $^{235}$U. In the frames shown all ions were extracted in the 2+ charge state from the chemically unpurified $^{233}$U source 1.




**Extended Data Figure 5 | Chart of nuclides potentially contained in the source material.** The chart includes all elements from curium (Cm, $Z=96$) to mercury (Hg, $Z=80$). All nuclides drawn are taken into consideration for the exclusion of a potential nuclear background. For completeness, all potentially populated nuclides are shown, even if their activity can be assumed to play a negligible role owing to a small branching ratio or a long half-life of the mother nuclide. These nuclides are shown without colour. Nuclides that can potentially recoil from the source as populated via $\alpha$ decay are assigned a white circle. Nuclides that possess one or more isomeric states carry a white star. A complete list of potentially contributing excited isomers is given in Extended Data Table 3. The short forms a, d and s are used for years, days and seconds. SF is short for spontaneous fission.




**Extended Data Figure 6 | $\alpha$-energy spectra of different Si-detector-based measurements, each accumulated for 7,200 s.** A silicon charged particle detector (Methods) is used for detection. The extracted ion species is chosen by mass-to-charge separation by the QMS. **a**–**d**, The accumulated counts are shown for extraction from the chemically unpurified $^{233}$U source 1 for $^{213}Bi^{2+}$ (**a**), no extraction (that is, dark counts, **b**), $^{229}Th^{2+}$ (**c**) and $^{229}Th^{3+}$ (**d**). No signal above the background is detected for $^{229}$Th in the 2+ and 3+ charge states. This clearly excludes any $\alpha$ decay as signal origin.

**Extended Data Table 1 | Potential background contributions and ways to exclude them**

| No. | Way of background exclusion | Type of background | | | |
|---|---|---|---|---|---|
| | | A | B | C | D |
| 1 | Signal comparison between $^{229}Th^{2+}$ and $^{233}U^{2+}$ | x | x | | |
| 2 | Comparative $^{229}Th^{2+}$ and $^{233}U^{2+}$ signal behaviour as a function of MCP surface voltage | x | x | | |
| 3 | Signal comparison between $^{229}Th$ and $^{230}Th$ | x | x | x | |
| 4 | Signal comparison between $^{229}Th^{2+}$ and $^{229}Th^{3+}$ | | | | x |
| 5 | Signal comparison between $^{229}Th$ originating from chemically purified and unpurified $^{233}U$ sources | | | | x |
| 6 | Exclusion based on signal appearance and the $^{229m}Th$ half-life limit | | | | x |
| 7 | Search for $\alpha$ and $\beta$ decays using Si and $LN_2$ cooled Si(Li) detectors | | | | x |

Column 1 lists the measurement numbers as given in the text, while in column 2 the corresponding measurement types are detailed. Each column from 3 to 6 corresponds to one type of potential background contribution. A cross indicates its exclusion by the measurement given in the corresponding row. Most of the potential background contributions could be excluded in multiple ways.
A, background from ionic kinetic energy or energy carried in the form of the charge state of the impinging ion. B, Background originating from the setup components ($^{233}U$ source, ion transport system, detection system). C, background from the thorium atomic shell. D, Background from activity other than $^{229}Th$.

**Extended Data Table 2 | Ionization energies of elements potentially contained in the $^{233}U$ source material**

| Element | Atomic no. | 1+ [eV] | 2+ [eV] | 3+ [eV] |
|---|---|---|---|---|
| Curium | 96 | 5.99 | 12.4 | 20.1 |
| Americium | 95 | 5.97 | 11.7 | 21.7 |
| Plutonium | 94 | 6.03 | 11.5 | 21.1 |
| Neptunium | 93 | 6.27 | 11.5 | 19.7 |
| Uranium | 92 | 6.19 | 11.6 | 19.8 |
| Protactinium | 91 | 5.89 | 11.9 | 18.6 |
| Thorium | 90 | 6.31 | 11.9 | 18.3 |
| Actinium | 89 | 5.38 | 11.8 | 17.4 |
| Radium | 88 | 5.28 | 10.1 | 31.0 |
| Francium | 87 | 4.07 | 22.4 | 33.5 |
| Radon | 86 | 10.75 | 21.4 | 29.4 |
| Astatine | 85 | 9.32 | 17.9 | 26.6 |
| Polonium | 84 | 8.41 | 19.3 | 27.3 |
| Bismuth | 83 | 7.29 | 16.7 | 25.6 |
| Lead | 82 | 7.42 | 15.0 | 31.9 |
| Thallium | 81 | 6.11 | 20.4 | 29.9 |
| Mercury | 80 | 10.44 | 18.7 | 34.5 |

From radium downwards, all elements reveal ionization potentials that are above the first ionization potential of helium ($E_{ion} = 24.6$ eV). Besides the mass-to-charge separation, this feature is also exploited to remove short-lived nuclides from the $^{229}Th^{3+}$ ion beam, as only elements with a third ionization potential below 24.6 eV can be extracted from the buffer-gas stopping cell to a significant extent in the 3+ charge state. For other elements, the 3+ charge state is reduced to the 2+ charge state during collisions with the helium buffer gas. Ionization energies from ref. 60.

**Extended Data Table 3 | Known isomeric states of nuclides potentially contained in the $^{233}$U source material**

| Isomer | Excitation energy | Half-life | Decay channel | Population |
|---|---|---|---|---|
| $^{244m}$Cm | 1.04 MeV | 34 ms | IT: 100.00 % | not populated |
| $^{242m1}$Am | 48.6 keV | 141 a | IT: 99.55 %, $\alpha$: 0.45 % | 100 % populated |
| $^{242m2}$Am | 2.20 MeV | 14.0 ms | SF: 100 %, $\alpha: \leq 5.0 \cdot 10^{-3}$ %, IT | not populated |
| $^{235m}$U | 76 eV | 26 min | IT: 100 % | 70 % from $^{239}$Pu |
| $^{234m}$Pa | 73.9 keV | 1.16 min | $\beta-$: 99.84 %, IT: 0.16 % | 78 % from $^{234}$Th |
| $^{229m}$Th | $\sim$7.8 eV | unknown | unknown | 2 % from $^{233}$U |
| $^{212m}$Po | 2.922 MeV | 45.1 s | $\alpha$: 99.93 %, IT: 0.07 % | not populated |
| $^{211m}$Po | 1.462 MeV | 25.2 s | $\alpha$: 99.98 %, IT: 0.02 % | not populated |
| $^{215m}$Bi | 1.348 MeV | 36.9 s | IT: 76.2 %, $\beta-$: 23.8 % | not populated |
| $^{212m1}$Bi | 0.250 MeV | 25.0 min | $\alpha$: 67.0 %, $\beta-$: 33.0 % | not populated |
| $^{212m2}$Bi | 1.91 MeV | 7.0 min | $\beta-$: 100 % | not populated |
| $^{210m}$Bi | 0.271 MeV | $3.04 \cdot 10^6$ a | $\alpha$: 100 % | not populated |
| $^{207m}$Pb | 1.633 MeV | 0.806 s | IT: 100 % | $8.1 \cdot 10^{-4}$ % from $^{211}$Po |
| $^{207m}$Tl | 1.348 MeV | 1.33 s | IT: 100 % | $9 \cdot 10^{-4}$ % from $^{211}$Bi |
| $^{206m}$Tl | 2.643 MeV | 3.74 min | IT: 100 % | not populated |

The isomeric excitation energies, half-lives, decay channels and population branching ratios are listed[1]. SF, spontaneous fission; IT, internal transition. The latter includes both means of de-excitation: by photon emission or by internal conversion.

doi:10.1038/nature17936

# Principles underlying sensory map topography in primary visual cortex

Jens Kremkow[1]*†, Jianzhong Jin[1]*, Yushi Wang[1] & Jose M. Alonso[1]

**The primary visual cortex contains a detailed map of the visual scene, which is represented according to multiple stimulus dimensions including spatial location, ocular dominance and stimulus orientation. The maps for spatial location and ocular dominance arise from the spatial arrangement of thalamic afferent axons in the cortex. However, the origins of the other maps remain unclear. Here we show that the cortical maps for orientation, direction and retinal disparity in the cat (*Felis catus*) are all strongly related to the organization of the map for spatial location of light (ON) and dark (OFF) stimuli, an organization that we show is OFF-dominated, OFF-centric and runs orthogonal to ocular dominance columns. Because this ON–OFF organization originates from the clustering of ON and OFF thalamic afferents in the visual cortex, we conclude that all main features of visual cortical topography, including orientation, direction and retinal disparity, follow a common organizing principle that arranges thalamic axons with similar retinotopy and ON–OFF polarity in neighbouring cortical regions.**

Orientation preference is systematically mapped as a pinwheel pattern in the primary visual cortex of primates and carnivores[1–3]. In this map, orientation changes rapidly around pinwheel centres and remains unchanged at the pinwheel blades. This organization is remarkably similar across these animals, suggesting a common organizing principle[4,5]; however, its anatomical substrate remains unknown. The anatomical substrate of orientation maps is unlikely to be determined by the structure of cortical neurons because cortical dendrites are not shaped by features of the orientation map[6] and rapid changes in orientation preference can occur within distances smaller than the diameter of a dendritic field[3]. Local intracortical connections among neurons with different orientation preferences could explain the broad orientation tuning near pinwheel centres, but recent results have indicated that these connections are biased towards neurons with similar orientations, even in animals such as the mouse that do not have orientation maps[7]. A possible anatomical substrate for orientation maps could be the axonal arrangement of ON and OFF thalamic afferents in the cortex[8–13], just as the substrate for ocular dominance maps is the arrangement of thalamic afferents from the contralateral and ipsilateral eyes[14]. Here, we provide support for this theory and conclude that thalamic afferents play a major role in shaping all topographic features of the primary visual cortex, including retinotopy, ocular dominance, orientation preference, direction preference and retinal disparity.

To study the relationship between changes in ON–OFF retinotopy and orientation preference, we introduced a multielectrode array horizontally into cat primary visual cortex (Fig. 1a) and targeted neurons within the middle cortical layers, which are the main recipients of thalamic inputs. We measured ON and OFF retinotopy with light and dark stimuli and used the ON–OFF difference to predict the preferred orientation of each cortical recording site (Fig. 1b; see also Extended Data Fig. 1a). Orientation tuning was measured with moving bars and represented as colour maps of response time-courses (Fig. 1c, left) and polar plots of response counts (Fig. 1c, right). The multielectrode recordings allowed us to study different regions of the cortical orientation map, including those containing abrupt changes in orientation preference (Fig. 1d, section from 1.5 mm to 1.7 mm) and direction preference (Fig. 1d, section from 0.1 mm to 0.2 mm).

## Cortical organization of ON–OFF retinotopy

Previous studies have shown that ON and OFF thalamic afferents are clustered in the visual cortex[15–18] but their spatial arrangement and relationship with other features of cortical topography are unknown. By measuring ON and OFF retinotopy along cortical horizontal penetrations, we show that ON and OFF cortical domains form interlaced patterns similar to ocular dominance patterns. Figure 2a illustrates a horizontal penetration crossing multiple interlaced ON and OFF domains. In this penetration, the retinotopy remained nearly constant at the peak of each domain and changed by about half a receptive field centre between domains of the same sign (for example, OFF to OFF).

The horizontal track illustrated in Fig. 2a ran roughly parallel to a single ocular dominance column for more than 2 mm. Figure 2b illustrates a different horizontal track that crossed ocular dominance columns perpendicularly (see also Extended Data Fig. 1b). As in the previous example, the retinotopy was nearly constant around the peak of each domain and changed by about half a receptive field centre between peaks of the same sign. However, unlike in Fig. 2a, the ON and OFF domains peaked at nearly the same cortical location (around the centre of the ocular dominance column). We did not find a pronounced mismatch in retinotopy between the two eyes at the borders of ocular dominance columns in cats, as has been reported in primates[14]. Instead, the retinotopy remained well matched in both spatial position and contrast polarity (Extended Data Figs 1b, 2). To quantify the topographic arrangement of ON and OFF domains, we calculated the correlation between normalized ON and OFF responses across cortical distance separately for penetrations that ran parallel or perpendicularly to ocular dominance columns (Fig. 2c; see Extended Data Fig. 1b and Methods for selection criteria). If the ON and OFF response strengths reached their maximum at different cortical locations (as in Fig. 2a), the correlation would approach a value of −1, whereas if they reached their maximum at the same cortical location (as in Fig. 2b), the correlation would approach a value of 1. The average correlation of the ON–OFF cortical

[1]Graduate Center for Vision Research, State University of New York, College of Optometry, 33 West 42nd Street, New York, New York 10036, USA. †Present address: Department of Biology, Institute for Theoretical Biology, Humboldt-Universität zu Berlin, Philippstrasse 13, 10115 Berlin, Germany.
*These authors contributed equally to this work.

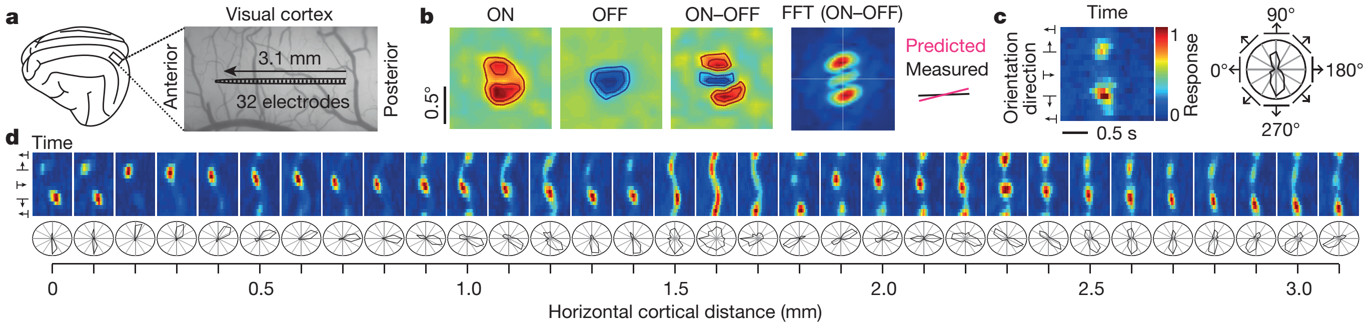



**Figure 1 | Recording from the horizontal dimension of visual cortex.** **a**, Recording configuration. **b**, Left, receptive fields mapped with light (ON) and dark (OFF) spots and ON–OFF receptive field difference. Right, orientation preference predicted by a 2D fast Fourier transform (FFT) from the ON–OFF receptive field difference. **c**, Orientation and direction tuning shown as response plot (left) and polar plot (right). **d**, Changes in orientation and direction preference across horizontal cortical distance.

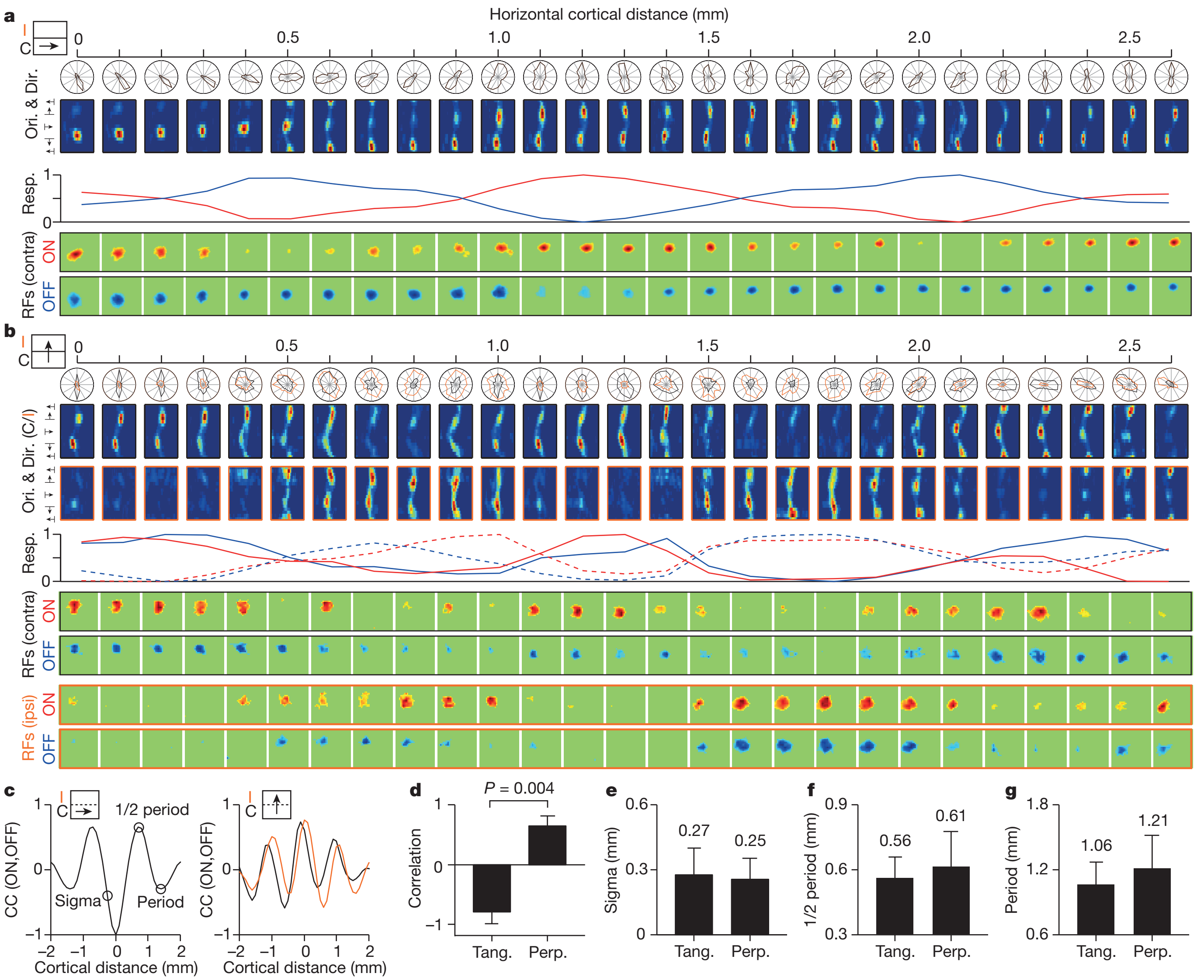



**Figure 2 | Topographic organization of ON and OFF cortical domains.** **a**, Example of a recording running parallel to an ocular dominance column. Icon on the left illustrates the recording (arrow) relative to the contralateral (C) and ipsilateral (I) columns. From top to bottom, the figure shows orientation tuning (polar and response plots), maximum ON (red) and OFF (blue) responses at each cortical site (line plot) and changes in ON and OFF receptive field position with cortical distance. **b**, Recording running perpendicular to ocular dominance columns (icon on the left) for contralateral (black) and ipsilateral (orange) eyes (continuous and dashed traces, respectively, in line plots). **c**, Cross-correlation between ON and OFF response profiles (red and blue lines, respectively, in **a** and **b**) in penetrations tangential (left) and perpendicular (right) to ocular dominance columns. **d**, Average correlation between ON and OFF response profiles in tangential (Tang.) penetrations ($n = 5$ penetrations, $n = 5$ animals) and perpendicular (Perp.) penetrations ($n = 6$ penetrations, $n = 4$ animals). **e**–**g**, Averages for spatial scale, half period and full period of ON–OFF correlation (average differences are not significant). All error bars are s.d. Statistical comparisons made with two-sided Wilcoxon tests.

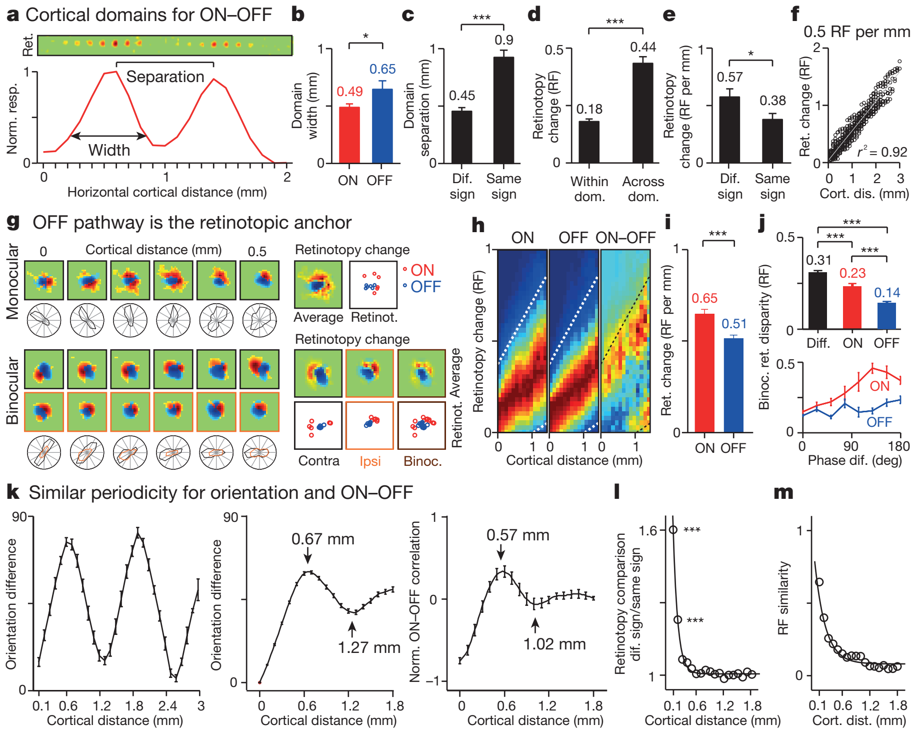



**Figure 3 | Cortical topographic relationships between ON–OFF, retinotopy and orientation preference.** **a**, Topography and retinotopy (Ret.) of two ON domains (receptive fields shown at top). **b**, OFF domains ($n = 20$ domains, $n = 12$ animals) are wider than ON domains ($n = 24$ domains, $n = 12$ animals). **c**, Domains of the same sign ($n = 16$ domains, $n = 12$ animals) are separated by twice as much distance as domains of different signs ($n = 31$ domains, $n = 12$ animals). **d**, Retinotopy changes more across domains (dom.) of the same sign ($n = 65$ domains, $n = 20$ animals) than within domains ($n = 125$ domains, $n = 20$ animals). **e**, Retinotopy changes more between domains of different (dif.) signs ($n = 31$ pairs of domains, $n = 12$ animals) than between domains of the same sign ($n = 16$ pairs of domains, $n = 12$ animals). **f**, Example recording showing smooth changes in retinotopy with cortical distance (Cort. dis.) at 0.5 receptive fields (RF) per mm ($n = 496$ paired comparisons). **g**, The OFF pathway anchors the cortical retinotopy of both monocular (top) and binocular (Binoc.) receptive fields (bottom; contralateral, black; ipsilateral, orange). ON responses (red) rotate around OFF responses (blue), as illustrated by individual series of receptive fields (left), receptive fields averaged across cortical distance (Average) and retinotopy of strongest ON and OFF responses (Retinot.). **h**, Retinotopy changes with cortical distance for ON, OFF and ON–OFF responses (red, maximum; blue, minimum). Dotted lines show 20% of maximum ON responses ($n = 2{,}603$ paired comparisons, $n = 8$ animals). **i**, Retinotopy changes are more restricted for OFF than ON responses ($n = 962$ ON and 962 OFF paired-comparisons, $n = 23$ animals). **j**, Binocular retinal disparity is smallest when measured between OFF subregions (top, $n = 502$ for ON–OFF, 251 for ON–ON and 251 for OFF–OFF subregions, $n = 28$ animals). ON retinal disparity changes more than OFF retinal disparity with differences in spatial phase (bottom). **k**, Periodicity in orientation preference across horizontal cortical distance within a single penetration (left) and across penetrations (middle, $n = 618$ paired comparisons, $n = 37$ animals). The orientation periodicity resembles the periodicity of the ON–OFF correlation (right; $n = 11$ penetrations, $n = 8$ animals). **l**, Retinotopy difference between subregions of different signs falls rapidly with cortical distance ($n = 13{,}416$ paired comparisons, $n = 23$ animals). **m**, Receptive field similarity also decays with cortical distance but at a slower rate ($n = 4{,}128$ paired comparisons, $n = 23$ animals). All error bars are s.e.m. $*P < 0.05$, $***P < 0.0001$ with two-sided Wilcoxon tests.

periodicity was $+0.65 \pm 0.17$ (mean ± s.d.) in penetrations perpendicular to ocular dominance columns and $-0.78 \pm 0.20$ in penetrations running tangentially (Fig. 2d, $P = 0.004$, Wilcoxon test), indicating that ON and OFF domains are interlaced along the main axis of the ocular dominance column but aligned along its perpendicular axis.

The periodicity of ON and OFF domains was similar in penetrations running tangentially and perpendicular to ocular dominance columns. It had a sigma of about 0.3 mm (Fig. 2e; $0.27 \pm 0.12$ mm and $0.25 \pm 0.10$ mm for tangential and perpendicular penetrations, respectively), a half period of about 0.6 mm (Fig. 2f; $0.56 \pm 0.10$ mm and $0.61 \pm 0.16$ mm) and a period of about 1.1 mm (Fig. 2g; $1.06 \pm 0.21$ mm and $1.21 \pm 0.31$ mm). To quantify in more detail the cortical spread and retinotopy change in each cortical domain, we selected penetrations that passed through a sequence of three or more ON and OFF domains (Fig. 3a; only ON domains shown for clarity). Consistent with our previous results[10], OFF cortical domains were significantly larger than ON cortical domains (Fig. 3b; OFF: $0.65 \pm 0.32$ mm, ON: $0.49 \pm 0.15$ mm, $P = 0.048$, Wilcoxon test) but were separated by similar cortical distances (ON to ON: $0.88 \pm 0.23$ mm; OFF to OFF: $1.0 \pm 0.24$ mm, $P = 0.3144$, Wilcoxon test), which was about twice the distance separating domains of different signs (Fig. 3c; $0.9 \pm 0.24$ mm versus $0.45 \pm 0.17$ mm, $P < 0.0001$, Wilcoxon test). The retinotopy change was limited to less than 0.2 receptive field centres within each domain and approached 0.5 receptive field centres between domains of the same sign (Fig. 3d; $0.18 \pm 0.12$ versus $0.44 \pm 0.24$ receptive field centres, $P < 0.0001$, Wilcoxon test). When normalized by cortical distance, the retinotopy moved faster between domains of different signs than domains of the same sign, probably because domains of different signs are less likely to share thalamic afferents (Fig. 3e; $0.57 \pm 0.39$ versus $0.38 \pm 0.21$ receptive field centres per mm, $P = 0.036$, Wilcoxon test).

## ON–OFF retinotopy and ocular dominance columns

Retinotopy is thought to change abruptly at the borders of ocular dominance columns in monkeys because of the interruption caused by the cortical representations of the two eyes[14]. Notably, our recordings revealed smooth changes in retinotopy in cats. To quantify these retinotopy changes, we selected tangential penetrations that passed through a sequence of at least three ocular dominance columns (for example, left–right–left, Extended Data Fig. 1b) and then measured how retinotopy changed between the peaks of ocular dominance columns for the same eye. Consistent with previous work[19], ocular dominance columns had an average width of around 0.5 mm in the cat

(0.44 ± 0.14 mm, $n = 31$) and ocular dominance columns for the same eye were separated from each other by around 1 mm (1.02 ± 0.17 mm, $n = 13$). Similar to the retinotopy changes between ON–OFF domains of the same sign (Fig. 3d), the retinotopy changes between ocular dominance columns of the same eye were about 0.5 receptive field centres (0.55 ± 0.22 receptive field centres, $n = 13$, data not shown). In fact, some cortical penetrations showed almost a perfect linear relationship between cortical distance and retinotopy with a slope of 0.5 receptive field centres per mm (Fig. 3f).

## OFF responses anchor cortical retinotopy

Our previous work demonstrated that OFF thalamic afferents cover larger cortical territory and make stronger connections than ON thalamic afferents in cat visual cortex[9,15]. Because of their larger horizontal extent, retinotopy should change less with cortical distance for OFF than ON cortical responses. We found not only that OFF retinotopy is more precise than ON retinotopy but also that it acts as the anchor of the cortical retinotopic map. This unexpected result, which we previously reported in an abstract[20], has now been replicated in tree shrew visual cortex[21] and it seems also to be present in primates (Extended Data Fig. 3). In horizontal penetrations through cat visual cortex, we frequently found that ON retinotopy rotated around OFF retinotopy (Fig. 3g), and that the retinotopy scatter was larger for ON than OFF responses (Fig. 3h–i; 0.65 ± 0.79 versus 0.51 ± 0.61 receptive field centres per mm, $P < 0.0001$, Wilcoxon test). Notably, OFF retinotopy anchored not only the monocular retinotopic map but also the binocular retinal disparity. In binocular receptive fields, the retinotopy changed less for OFF than ON responses and, although OFF retinotopy tended to be spatially aligned between the two eyes, ON retinotopy rotated around OFF (Fig. 3g, bottom; see also Extended Data Fig. 3 for an example in a macaque). Binocular retinal disparity was largest for receptive field subregions of different signs, intermediate for ON–ON subregions and smallest for OFF–OFF subregions (Fig. 3j, top; 0.31 ± 0.18, 0.23 ± 0.20 and 0.14 ± 0.11 receptive field

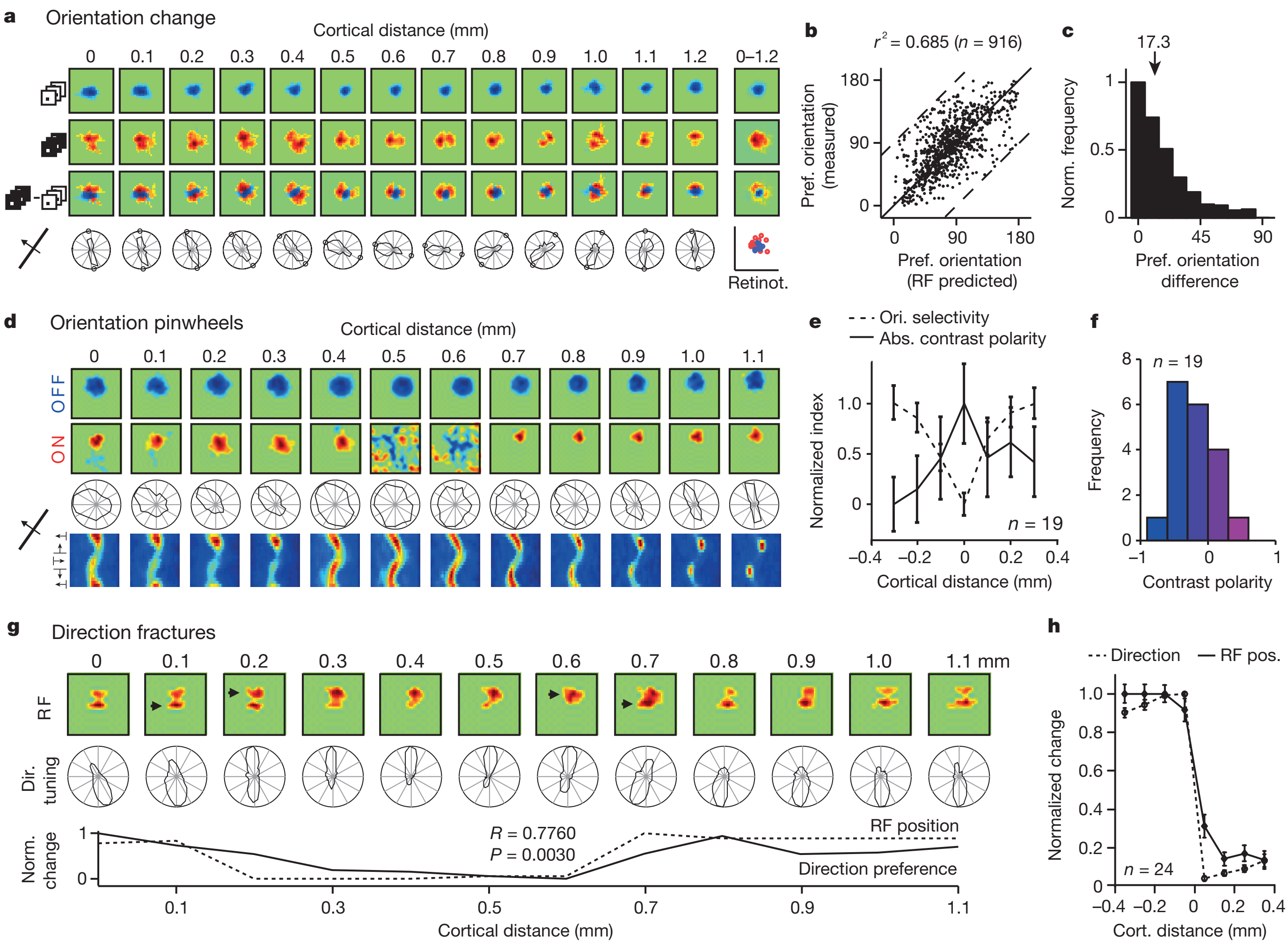



**Figure 4 | Changes in retinotopy explain changes in orientation and direction preference throughout the cortex. a**, Horizontal penetration showing a strong relationship between changes in ON–OFF retinotopy and orientation preference. Responses to light stimuli (middle) rotate around responses to dark stimuli (top) as seen in the dark–light difference (bottom). Orientation and direction tuning and ON/OFF retinotopy are shown below the colour panels (small circles in polar plots are orientation predictions based on dark–light receptive fields). **b**, Predicted and measured comparisons in 109 penetrations (916 recording sites, $n = 26$ animals) that passed our selection criteria (see Methods; dashed lines mark maximum possible mismatch). **c**, Normalized count of differences between measurements and predictions (median, 17.3°). **d**, Horizontal penetration passing through a pinwheel (at 0.5–0.6 mm) that was completely OFF dominated. **e**, Pinwheel centres (aligned at cortical distance zero) tended to have higher absolute (Abs.) contrast polarity (strong OFF or ON dominance) than their cortical neighbourhoods ($n = 19$ penetrations, $n = 13$ animals; $P < 0.0001$ for difference in orientation (Ori.) selectivity and $P = 0.039$ for difference in absolute contrast polarity when comparing 0 and ±0.3 mm, one-sided Wilcoxon tests). **f**, Histogram showing the contrast polarity of the 19 pinwheels from **e**. **g**, Horizontal penetration passing through regions with abrupt changes in direction preference (0.1–0.3 mm and 0.6–0.7 mm). Abrupt changes in direction were associated with abrupt changes in retinotopy (arrows at top and line plots at bottom). **h**, Aligning direction reversals at cortical distance zero ($n = 24$ penetration sections, $n = 10$ animals) revealed a strong association between direction and retinotopy changes (RF pos). All error bars are s.e.m.

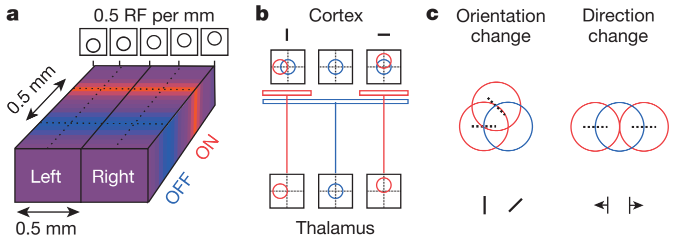



**Figure 5 | Principles underlying sensory map topography in primary visual cortex. a**, ON and OFF domains run perpendicular to ocular dominance columns and are separated by ~0.5 mm from each other. Retinotopy changes smoothly at ~0.5 receptive fields per mm. **b**, Schematic showing how thalamo-cortical architecture could make ON receptive fields rotate around OFF receptive fields. **c**, Diagram explaining how changes in ON–OFF retinotopy result in changes in orientation and direction preference.

centres, respectively; $P < 0.0001$, Wilcoxon test). Moreover, OFF binocular retinal disparity remained small even if differences in relative spatial phase increased, whereas ON retinal disparity could change by nearly 0.5 receptive field centres (Fig. 3j, bottom). These results indicate that retinotopy is matched at the borders of ocular dominance columns not only in spatial position but also in ON–OFF contrast polarity. This binocular match in ON–OFF retinotopy is not very different from that observed in the mouse[22], an animal that does not have ocular dominance columns or orientation maps. However, the ON–OFF retinotopic match in the cat is most precise for OFF cortical responses, which act as the anchor of both monocular retinotopy and binocular retinal disparity. The limited retinotopy changes at the borders of ocular dominance columns seem ideal to generate a smooth and precise map of retinal disparity[23].

## ON–OFF retinotopy and orientation preference

Our previous work showed that the arrangement of OFF and ON thalamic afferents in the visual cortex is closely related to the representation of orientation preference[9]. To quantify this relationship across the horizontal dimension of the visual cortex, we first compared the average periodicity of ON–OFF retinotopy with the periodicity of orientation preference across cortical distance. The periodicity of orientation preference was very pronounced even in single horizontal penetrations (Fig. 3k, left and Extended Data Fig. 4) and, on average, it had a half period of 0.67 mm and a full period of 1.27 mm (Fig. 3k, middle), which closely matched the average periodicity of ON–OFF retinotopy (Fig. 3k, right; average periodicity of ON–OFF retinotopy 0.57/1.02 mm; 0.56/1.06 mm for tangential penetrations and 0.61/1.21 mm for perpendicular penetrations). The difference in retinotopy between neurons separated by 0.1 mm was 1.6-times larger for subregions of different signs (ON–OFF) than for subregions of the same sign (Fig. 3l). However, the different/same-sign ratio decayed rapidly with cortical distance to 89% at 0.3 mm (Fig. 3l) and 0.3 mm is the approximate size of a cortical orientation domain in cats[24]. Receptive field similarity (as defined by correlation coefficient) also decayed with cortical distance but at a much slower rate (Fig. 3m; 87% at 1 mm).

The relationship between ON–OFF retinotopy and orientation preference was pronounced when we selected horizontal penetrations that passed through cortical regions with marked ON–OFF spatial segregation and good orientation selectivity (Fig. 4a and Extended Data Figs 5, 6). In these penetrations, the orientation preference measured with moving bars was strongly correlated with the orientation preference predicted from the ON–OFF receptive field structure (Fig. 4b; $r^2 = 0.68$, $P < 0.0001$; median $r^2$ within-penetration: 0.75; see Methods for selection criteria) and the median prediction error was only 17.3° (Fig. 4c; probability that the distribution is uniform random: $P < 0.0001$, Wilcoxon test). The predictions of orientation preference were not as good in horizontal cortical penetrations that had receptive fields strongly dominated by one contrast polarity, as our methods were not sensitive enough to measure the retinotopy of weak, non-dominant responses. In particular, pinwheel centres had a tendency to be more dominated by one contrast polarity than adjacent cortical regions (Fig. 4d, e) and most of them were OFF dominated (Fig. 4f; pinwheel defined as monocular recording site responding to all stimulus orientations). This result is consistent with the notion that OFF thalamic afferents cover more cortical space and make stronger connections than ON thalamic afferents[9,15]. It should be noted, however, that few pinwheels were completely OFF dominated (Fig. 4d shows one example), and none was completely ON dominated (Fig. 4f). The lack of purely OFF dominated or ON dominated pinwheels is consistent with the spread of thalamic axons, which can be more than 1 mm along the main axis of an ocular dominance column[25], whereas the average separation between ON and OFF domains is only 0.5 mm (Fig. 3c). Also consistent with the OFF dominance of visual cortex, regions in which ON retinotopy rotated around OFF ($n = 15$ regions) were more frequent than regions in which OFF retinotopy rotated around ON ($n = 4$ regions; Extended Data Fig. 7).

## ON–OFF retinotopy and direction preference

Although our previous work predicted that cortical changes in ON–OFF retinotopy should be related to changes in orientation preference[9], we were surprised to find that changes in ON–OFF retinotopy were also related to changes in direction preference. Because weaker receptive field subregions generate responses with longer response latencies than those of stronger subregions, cortical responses coincide in time and reinforce each other when a stimulus moves from a weak to a strong subregion but not from a strong to a weak subregion[26–31]. In cortical horizontal penetrations that passed through direction fractures (rapid reversals of direction preference), abrupt changes in the retinotopic position of the strongest receptive field subregion were associated with abrupt changes in direction preference (Fig. 4g). To quantify this relationship more carefully, we selected penetrations in which direction preference changed abruptly but orientation remained relatively constant (to avoid rotations or translations in retinotopy that were not related to direction). In 24 penetrations that met this criterion, rapid reversals in direction preference (Fig. 4h, marked as 0 cortical distance) were strongly associated with rapid changes in the retinotopy of the strongest receptive field subregion and both occurred within 0.1 mm of each other.

## Discussion

Our findings suggest that the topography of the visual cortex in carnivores and primates is governed by a precise match in the properties of the thalamic afferents that converge at a given cortical point. The afferents are precisely matched in retinotopy, which changes slowly at 0.5 receptive field centres per mm in cats (Fig. 5a). They are also matched in eye input and ON–OFF polarity, which leads to a columnar organization for both ocular dominance[14] and ON–OFF responses[10,32] (Fig. 5a). In OFF domains, which are most prominent, OFF afferents are better matched in retinotopy than ON afferents; the opposite is true in ON domains. In this OFF-dominated and OFF-centric topography, changes in orientation and direction preference are determined by changes in ON–OFF retinotopy. Therefore, orientation preference may show a tendency to remain constant across the border of ocular dominance columns[33] simply because ON–OFF retinotopy also remains constant (Fig. 5a).

It is unclear what developmental mechanisms could generate this precise ON–OFF retinotopic match at each cortical point. However, if OFF domains with precisely matched retinotopy appear first during development[34], the retinotopy of the ON afferents may have to be displaced within each OFF domain so that ON and OFF afferents can simultaneously drive the same cortical targets (Fig. 5b). This mechanism would make ON receptive fields rotate around OFF receptive fields and, as a consequence, orientation and direction maps would originate (Fig. 5c) in a sensory map that is represented as continuously

as possible[14,35]. In the visual cortex, this continuous representation could be accomplished by precisely matching the response properties of ON and OFF thalamic afferents; however, the same principles may apply to other sensory spaces and afferents feeding other cortical areas that have maps for touch, hearing or spatial navigation[36–40].

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 11 November 2015; accepted 21 March 2016.**

**Published online 27 April 2016.**

1. Blasdel, G. G. & Salama, G. Voltage-sensitive dyes reveal a modular organization in monkey striate cortex. *Nature* **321,** 579–585 (1986).
2. Bonhoeffer, T. & Grinvald, A. Iso-orientation domains in cat visual cortex are arranged in pinwheel-like patterns. *Nature* **353,** 429–431 (1991).
3. Ohki, K. *et al.* Highly ordered arrangement of single neurons in orientation pinwheels. *Nature* **442,** 925–928 (2006).
4. Kaschube, M. *et al.* Universality in the evolution of orientation columns in the visual cortex. *Science* **330,** 1113–1116 (2010).
5. Nauhaus, I. & Nielsen, K. J. Building maps from maps in primary visual cortex. *Curr. Opin. Neurobiol.* **24,** 1–6 (2014).
6. Levy, M., Lu, Z., Dion, G. & Kara, P. The shape of dendritic arbors in different functional domains of the cortical orientation map. *J. Neurosci.* **34,** 3231–3236 (2014).
7. Cossell, L. *et al.* Functional organization of excitatory synaptic strength in primary visual cortex. *Nature* **518,** 399–403 (2015).
8. Miller, K. D. A model for the development of simple cell receptive fields and the ordered arrangement of orientation columns through activity-dependent competition between ON- and OFF-center inputs. *J. Neurosci.* **14,** 409–441 (1994).
9. Jin, J., Wang, Y., Swadlow, H. A. & Alonso, J. M. Population receptive fields of ON and OFF thalamic inputs to an orientation column in visual cortex. *Nature Neurosci.* **14,** 232–238 (2011).
10. Wang, Y. *et al.* Columnar organization of spatial phase in visual cortex. *Nature Neurosci.* **18,** 97–103 (2015).
11. Chapman, B. & Godecke, I. Cortical cell orientation selectivity fails to develop in the absence of ON-center retinal ganglion cell activity. *J. Neurosci.* **20,** 1922–1930 (2000).
12. Chapman, B., Zahs, K. R. & Stryker, M. P. Relation of cortical cell orientation selectivity to alignment of receptive fields of the geniculocortical afferents that arborize within a single orientation column in ferret visual cortex. *J. Neurosci.* **11,** 1347–1358 (1991).
13. Paik, S. B. & Ringach, D. L. Retinal origin of orientation maps in visual cortex. *Nature Neurosci.* **14,** 919–925 (2011).
14. Hubel, D. H. & Wiesel, T. N. Ferrier lecture. Functional architecture of macaque monkey visual cortex. *Proc. R. Soc. Lond. B Biol. Sci.* **198,** 1–59 (1977).
15. Jin, J. Z. *et al.* On and off domains of geniculate afferents in cat primary visual cortex. *Nature Neurosci.* **11,** 88–94 (2008).
16. Zahs, K. R. & Stryker, M. P. Segregation of ON and OFF afferents to ferret visual cortex. *J. Neurophysiol.* **59,** 1410–1429 (1988).
17. McConnell, S. K. & LeVay, S. Segregation of on- and off-center afferents in mink visual cortex. *Proc. Natl Acad. Sci. USA* **81,** 1590–1593 (1984).
18. Norton, T. T., Rager, G. & Kretz, R. ON and OFF regions in layer IV of striate cortex. *Brain Res.* **327,** 319–323 (1985).
19. Kaschube, M. *et al.* The pattern of ocular dominance columns in cat primary visual cortex: intra- and interindividual variability of column spacing and its dependence on genetic background. *Eur. J. Neurosci.* **18,** 3251–3266 (2003).
20. Kremkow, J. *et al.* Asymmetries in ON and OFF cortical retinotopy: are OFF receptive fields the anchors of cortical retinotopic maps? *Soc. Neurosci.* abstr. 639.09 (2013).
21. Lee, K.-S., Huang, X. & Fitzpatrick, D. Specificity in the spatial organization of receptive fields supporting multiple functional maps in tree shrew visual cortex *Soc. Neurosci.* abstr. 232.13 (2015).
22. Sarnaik, R., Wang, B. S. & Cang, J. Experience-dependent and independent binocular correspondence of receptive field subregions in mouse visual cortex. *Cereb. Cortex* **24,** 1658–1670 (2014).
23. Kara, P. & Boyd, J. D. A micro-architecture for binocular disparity and ocular dominance in visual cortex. *Nature* **458,** 627–631 (2009).
24. Sharma, J., Angelucci, A. & Sur, M. Induction of visual orientation modules in auditory cortex. *Nature* **404,** 841–847 (2000).
25. Humphrey, A. L., Sur, M., Uhlrich, D. J. & Sherman, S. M. Projection patterns of individual X- and Y-cell axons from the lateral geniculate nucleus to cortical area 17 in the cat. *J. Comp. Neurol.* **233,** 159–189 (1985).
26. Reid, R. C., Soodak, R. E. & Shapley, R. M. Linear mechanisms of directional selectivity in simple cells of cat striate cortex. *Proc. Natl Acad. Sci. USA* **84,** 8740–8744 (1987).
27. Jagadeesh, B., Wheat, H. S. & Ferster, D. Linearity of summation of synaptic potentials underlying direction selectivity in simple cells of the cat visual cortex. *Science* **262,** 1901–1904 (1993).
28. Tolhurst, D. J. & Dean, A. F. Evaluation of a linear model of directional selectivity in simple cells of the cat's striate cortex. *Vis. Neurosci.* **6,** 421–428 (1991).
29. Albrecht, D. G. & Geisler, W. S. Motion selectivity and the contrast-response function of simple cells in the visual cortex. *Vis. Neurosci.* **7,** 531–546 (1991).
30. McLean, J., Raab, S. & Palmer, L. A. Contribution of linear mechanisms to the specification of local motion by simple cells in areas 17 and 18 of the cat. *Vis. Neurosci.* **11,** 271–294 (1994).
31. Livingstone, M. S. Mechanisms of direction selectivity in macaque V1. *Neuron* **20,** 509–526 (1998).
32. Smith, G. B., Whitney, D. E. & Fitzpatrick, D. Modular representation of luminance polarity in the superficial layers of primary visual cortex. *Neuron* **88,** 805–818 (2015).
33. Blasdel, G. G. Orientation selectivity, preference, and continuity in monkey striate cortex. *J. Neurosci.* **12,** 3139–3161 (1992).
34. Albus, K. & Wolf, W. Early post-natal development of neuronal function in the kitten's visual cortex: a laminar analysis. *J. Physiol. (Lond.)* **348,** 153–185 (1984).
35. Swindale, N. V., Shoham, D., Grinvald, A., Bonhoeffer, T. & Hubener, M. Visual cortex maps are optimized for uniform coverage. *Nature Neurosci.* **3,** 822–826 (2000).
36. Woolsey, T. A. & Van der Loos, H. The structural organization of layer IV in the somatosensory region (SI) of mouse cerebral cortex. The description of a cortical field composed of discrete cytoarchitectonic units. *Brain Res.* **17,** 205–242 (1970).
37. Friedman, R. M., Chen, L. M. & Roe, A. W. Modality maps within primate somatosensory cortex. *Proc. Natl Acad. Sci. USA* **101,** 12724–12729 (2004).
38. Miller, L. M., Escabi, M. A., Read, H. L. & Schreiner, C. E. Functional convergence of response properties in the auditory thalamocortical system. *Neuron* **32,** 151–160 (2001).
39. Hafting, T., Fyhn, M., Molden, S., Moser, M. B. & Moser, E. I. Microstructure of a spatial map in the entorhinal cortex. *Nature* **436,** 801–806 (2005).
40. Peyrache, A., Lacroix, M. M., Petersen, P. C. & Buzsaki, G. Internally organized mechanisms of the head direction sense. *Nature Neurosci.* **18,** 569–575 (2015).

**Acknowledgements** We thank R. Lashgari, S. Komban, M. Jansen, C. Pons and E. Koch for helping collect data in some experiments; H. Swadlow, Q. Zaidi and R. Mazade for comments on the manuscript; and A. Movshon for lending some experimental equipment. This work was supported by the US National Institutes of Health (EY005253, J.M.A.), a DFG Research Fellowship (KR 4062/1-1, J.K.) and Humboldt-Universität zu Berlin in the framework of the Excellence Initiative of the BMBF and DFG (J.K.).

**Author Contributions** J.K., J.J., Y.W. and J.M.A. conducted the experiments and data analysis. J.K., J.J. and J.M.A. wrote the manuscript.

**Author Information** Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.M.A. (jalonso@sunyopt.edu).

## METHODS

All procedures were performed in accordance with the guidelines of the US Department of Agriculture and approved by the Institutional Animal Care and Use Committee at the State University of New York, State College of Optometry.

**Surgery and preparation.** Adult male cats (aged 6–12 months, $n = 40$) were tranquilized with acepromazine ($0.2\,\mathrm{mg\,kg^{-1}}$, intramuscularly) and initially anaesthetized with ketamine ($10\,\mathrm{mg\,kg^{-1}}$, intramuscularly). An intravenous catheter was inserted into each hind limb to allow continuous infusions of propofol ($5–6\,\mathrm{mg\,kg^{-1}\,h^{-1}}$) and sufentanil ($10–20\,\mathrm{ng\,kg^{-1}\,h^{-1}}$) for anaesthesia, vecuronium bromide ($0.2\,\mathrm{mg\,kg^{-1}\,h^{-1}}$) for muscle paralysis, and saline ($1–3\,\mathrm{ml\,h^{-1}}$) for hydration. All vital signs were closely monitored and carefully maintained within normal physiological limits. The nictitating membranes were retracted with 2% neosynephrine and the pupils dilated with 1% atropine sulphate. Contact lenses were used to protect the corneas and focus visual stimuli on the retina. The positions of the optic disc and the area centralis were plotted on a screen in front of the animal using a fibre optic light source. Details of the surgical procedures have been described previously[10]. We also performed recordings in one male rhesus macaque (age, 8.5 years; 10 kg) using similar procedures to those described above. The macaque was anaesthetized with ketamine ($10\,\mathrm{mg\,kg^{-1}}$, intramuscularly) and diazepam ($0.75\,\mathrm{mg\,kg^{-1}}$, intravenous) followed by propofol ($1.8\,\mathrm{mg\,kg^{-1}}$, intravenous) and a continuous infusion of sufentanil citrate that was maintained throughout the experiment ($6–20\,\mathrm{\mu g\,kg^{-1}\,h^{-1}}$, intravenous). The animal was paralysed after finishing the surgery with vecuronium bromide ($0.1\,\mathrm{mg\,kg^{-1}\,h^{-1}}$, intravenous).

**Electrophysiological recordings and data acquisition.** We used linear 32-channel multielectrode arrays (inter-electrode distance, 0.1 mm; Neuronexus) to record multi-unit neuronal activity along the horizontal dimension of primary visual cortex (Fig. 1a). The signals from the recording electrodes were amplified, filtered, and collected by a computer running Rasputin (Plexon), as previously described[10]. The multielectrode arrays were introduced with a small angle nearly parallel to the cortical surface (<5°), parallel to the anteroposterior axis in the middle of the posterolateral gyrus and centred in layer 4. The centring of the recordings in layer 4 was estimated from cortical depth, local field potentials and the presence of simple receptive fields measured with white noise, which are mostly restricted to layers 4 and 6 in cat visual cortex[10,41]. Sample size was chosen to be the largest possible for each analysis performed. All comparisons were evaluated for statistical significance using two-sided Wilcoxon tests (signed-rank for paired data and rank-sum for non-paired), except for that shown in Fig. 4e (one-sided Wilcoxon test). Data distributions are described in the main text by their mean and s.d. (median for Fig. 4c) while the figures show either s.d. or s.e.m. (see figure legends). No randomization was used to determine how samples or animals were allocated to experimental groups and no blinding approach was used for sample selection.

**Visual stimulation.** Visual stimuli were generated in Matlab (The MathWorks) using the Psychophysics Toolbox extensions[42] and presented on a calibrated CRT monitor (refresh rate 120 Hz, mean luminance $61\,\mathrm{cd\,m^{-2}}$). The monitor was positioned so that the receptive fields of all recorded channels were covered by the visual stimulus. We used light and dark moving bars (16 directions, 8 orientations) to measure orientation tuning (Fig. 1c) and receptive fields were mapped using sparse noise stimuli. The frames of the sparse noise were updated at a rate of 30 Hz (monitor refresh rate 120 Hz) and the sparse noise targets were either light ($120\,\mathrm{cd\,m^{-2}}$) or dark ($<2\,\mathrm{cd\,m^{-2}}$). Light targets were presented on a dark background and dark targets on a light background (Extended Data Fig. 1a). We used large targets (1–2° width) to drive responses from weak receptive field flanks. The use of large stimuli greatly overestimates the size of the receptive fields but provides a reliable estimate of the receptive field centre of mass (retinotopy). Visual stimuli were presented to one eye at a time (monocular stimulation).

**Data analysis.** All data analysis was performed in Matlab using customized analysis routines as described below for each major set of measurements.

**Orientation selectivity and receptive field analysis.** Orientation tuning was measured with moving bars (16 directions of motion) and fitted with a von Mises function[43]. The orientation or direction preference and selectivity were extracted from the fits as previously described[44]. To precisely estimate the spatial ON and OFF receptive fields of each recording site, we calculated the peri-stimulus-time histogram (PSTH) at a temporal resolution of 1 ms for each stimulus pixel. This analysis resulted in a 3D array ($x$-space, $y$-space, time) representing the neuronal response in space and time. We then estimated the spatial receptive field by integrating all spikes caused by the stimulus onset (Extended Data Fig. 1a, grey shaded area in the PSTH) after smoothing the temporal response with a Gaussian window (sigma, 10 ms). The ON receptive fields were calculated from the response onset to light targets and the OFF receptive fields from the onset to dark targets. This analysis resulted in four receptive field measurements for each cortical site (contra eye: ONc and OFFc; ipsi eye: ONi and OFFi). Each receptive field was then normalized by subtracting its mean and dividing by its maximum. The normalized ON and OFF receptive fields were then used to calculate the ON–OFF receptive fields by subtracting OFF from ON. When showing receptive fields to compare changes in ON–OFF retinotopy across the cortex (Fig. 2a), we normalized by the maximum response to ON or OFF, whichever was greater (normalization for contrast polarity). When showing binocular receptive fields to compare changes in ocular dominance (Fig. 2b), we normalized by the maximum response of both eyes, whichever was greater (normalization for ocular dominance). The receptive field integration time was 50 ms to measure ON–OFF retinotopy (Fig. 2), 200 ms to measure contralateral/ipsilateral retinotopy (Extended Data Fig. 1b) and variable to predict orientation preference (same procedure as explained in 'binocular organization of ON–OFF' below). Receptive field similarity across recording sites was estimated by calculating the correlation coefficient between the ON–OFF receptive fields.

**Binocular alignment of receptive fields.** To measure the binocular organization of ON and OFF retinotopy, we first had to align the monocular receptive fields because the eyes were misaligned by the muscle paralysis in our preparation. To achieve unbiased eye alignment, we made use of the high number of simultaneously measured receptive fields (32 recording positions), using an approach that was very successful at revealing cortical maps for retinal disparity[23]. To that end, we calculated the retinotopic receptive field (Rr) by summing the ON and OFF receptive fields of all channels, separately for the ipsilateral ($\mathrm{Rr_i}$) and contralateral eye ($\mathrm{Rr_c}$). We then performed a 2D cross-correlation analysis between $\mathrm{Rr_i}$ and $\mathrm{Rr_c}$ to estimate the horizontal and vertical shift between the two eyes and used this measurement to align both eyes.

**Cortical domains for ON–OFF.** To calculate the cortical ON–OFF domains, we analysed the neuronal responses to light and dark sparse noise stimuli. For each cortical site we calculated the spatial receptive fields (ONc, ONi, OFFc, OFFi) at the peak of the response onset (the temporal response was smoothed with a Gaussian window; sigma, 10 ms). To extract the relative strength between ON–OFF and ipsilateral–contralateral responses, we normalized the amplitude of the receptive fields by the strongest response at each cortical site. A small Gaussian window (sigma, 1 recording channel) was used to smooth the responses across cortex. This analysis resulted in a 3D array for each stimulus condition (ONc, ONi, OFFc, OFFi), representing $x$ and $y$ of the visual field (retinotopy) and the 32 recording channels (cortical distance). From this 3D representation of the ON–OFF cortical domains, we calculated the 1D cortical activation profiles (Fig. 2a, b; red and blue traces) by using the value of the maximum response at each cortical site. This analysis resulted in 1D activation profiles for ONc, ONi, OFFc and OFFi that represented the relative strength of ON–OFF and ipsilateral–contralateral responses at each cortical position. To estimate the correlation, spatial scale and periodicity of the ON–OFF responses across cortical distance, we calculated the cross-correlation between the ON and OFF cortical activation profiles (Fig. 2a, b; red and blue traces). We used the correlation coefficient between ON and OFF as the measure for the overall correlation between ON–OFF domains. The spatial spread was estimated as the standard deviation of a Gaussian function fitted to the central part of the cross-correlogram (Fig. 2c). The half period was taken as the first reversal in the cross-correlogram and the full period as the second peak (Fig. 2c). To compare the cortical widths of ON and OFF domains, we selected horizontal cortical penetrations that crossed at least three ON or OFF cortical domains (Fig. 3a). We then measured the width of each domain as the number of contiguous recording sites that generated responses with high signal-to-noise ratio (SNR > 5) and averaged the widths separately for ON and OFF domains (Fig. 3b). The cortical distance between domains was measured between the most central recording sites within each domain (Fig. 3c). The retinotopy change was measured as the difference in retinotopy between two recording sites, using the larger receptive field diameter as the unit (Fig. 3d, e).

To compare the ON–OFF arrangement to ocular dominance columns, we selected our longest horizontal recording tracks that either remained monocular for the same eye or alternated between monocular responses for left and right eyes along the track length. We assumed that a horizontal track that remained monocular for the same eye for more than 1.2 mm was running roughly tangentially to an ocular dominance column and that a track that showed multiple alternating monocular responses for left and right eyes was running roughly perpendicular. Following this strict criteria, five horizontal tracks were classified as tangential to an ocular dominance column (average track length and range: $1.74 \pm 0.5$ mm, 1.2–2.6 mm; average and range of ON/OFF domain number: $3.2 \pm 0.97$, 2–5) and six tracks were classified as perpendicular (average track length and range: $2.23 \pm 0.51$ mm, 1.4–2.9 mm; average and range of ocular domain number: $4 \pm 1.52$, 2–6).

**Cortical domains for ocular dominance.** We selected horizontal cortical penetrations that passed through at least three different ocular dominance domains. The width of each domain was measured as the number of contiguous recording sites that generated responses with high SNR (SNR > 5). The cortical distance between the peaks of ocular dominance domains was measured as the distance between the

central recording sites within each domain. The retinotopy change was measured as the difference in retinotopy between receptive fields located between the peaks of ocular dominance columns for different eyes using the larger receptive field diameter as the unit.

**Retinotopy change of ON–OFF responses.** To estimate the retinotopy change across the horizontal dimension of cortex, we measured the centre of the strongest receptive field subregion by calculating the centre of mass around the peak response (using a receptive field threshold at 70–80% of maximum response). We then calculated the Euclidian distances between the receptive field centres of paired recording sites and normalized this distance by the diameter of the larger receptive field. The receptive field diameter was approximated from the area of the receptive field with a response above 20% of the maximum response (assuming a circular receptive field). To maximize the accuracy of our measurements, the population analysis included only cortical sites with SNR $>$ 10.

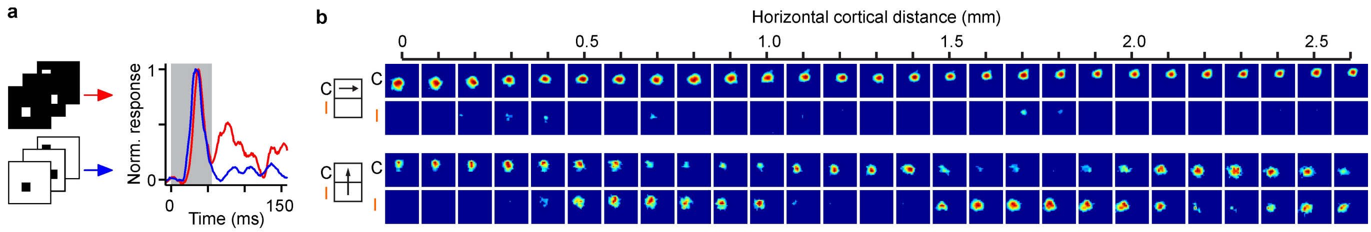
**Binocular organization of ON–OFF retinotopy.** To study the binocular organization of the ON–OFF retinotopy, we fitted a 2D Gabor function to the ON–OFF receptive fields (ONc–OFFc, ONi–OFFi). We then extracted the spatial phase difference from the Gabor fits and measured binocular disparity as the retinotopic distance between the positions of the subregions from the ON–OFF receptive fields. We calculated the ON–OFF receptive field by optimizing ON–OFF segregation, as this resulted in better and more reliable fits to the 2D Gabor function. To achieve this, we used a sliding window of 50 ms and calculated the ON–OFF receptive field with a range of starting positions (0–100 ms). From this ensemble of ON–OFF receptive fields, we selected the one that had the highest SNR and most balanced ON–OFF receptive field. ON–OFF balance was calculated as the absolute value of contrast polarity, where contrast polarity is (max(ON) $-$ max(OFF))/(max(ON) $+$ max(OFF)). If the absolute contrast polarity equals 0, ON and OFF responses are equally strong; if it equals 1, responses are completely dominated by either OFF or ON. Because the spatial phase can vary over the time course of the spatiotemporal receptive field[45], we always used the same time point to calculate the ON–OFF receptive fields in both eyes. To maximize the accuracy of the measurements, the population analysis included only sites with ON–OFF receptive fields that had SNR $>$ 6 and were well fit by the Gabor function (goodness of fit $>$ 0.5).

**Orientation and ON–OFF periodicity.** To study the orientation periodicity, we extracted the orientation preference from the fitted tuning curves (see above) and then calculated the orientation difference as a function of cortical distance. We measured the orientation difference between all possible pairs on our 32-channel recording array ($n = 496$ per recording array). We repeated this analysis across our entire data set and calculated the median orientation difference for each cortical distance (Fig. 3k, middle). To ensure that the measurement was precise, we included only pairs with excellent fits in orientation tuning (goodness of fit $>$ 0.9), pronounced orientation selectivity (orientation selectivity index $>$ 0.5) and responses with high SNR (SNR $>$ 4), resulting in 20,672 pairs across all possible cortical distances (orientation selectivity was defined as the ratio between the response at the preferred orientation and the response at the orthogonal orientation). We then estimated the half period from the first reversal of the average orientation difference across cortical distance and the full period from the second minimum (Fig. 3k, middle). To characterize the periodicity of ON–OFF responses across cortical distance, we averaged all cross-correlation measurements from ON–OFF cortical domains (Fig. 2c). Because ON–OFF domains are anticorrelated in recordings tangential to the ocular dominance bands but correlated in recordings perpendicular to ocular dominance bands (Fig. 2d), we multiplied the cross-correlograms of the recordings perpendicular to ocular dominance columns by $-1$ before averaging. We then obtained periodicity measures from the average normalized ON–OFF correlation for both the half period and the full period (Fig. 3k, right).

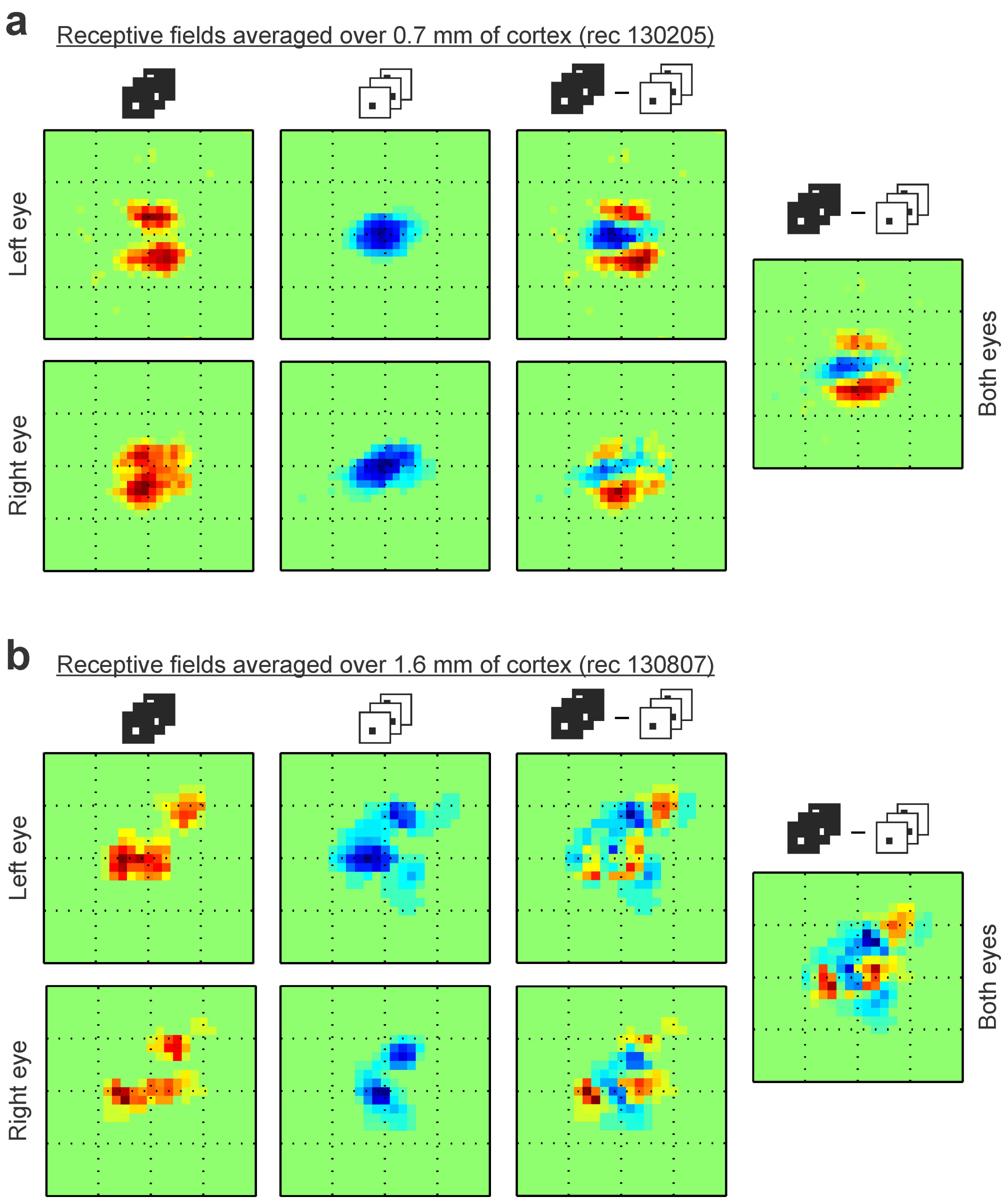
**Predicting orientation preference from the receptive field.** To predict the orientation preference from the ON–OFF receptive fields, we first calculated the ON–OFF receptive field difference using the sliding window approach described above (see Binocular organization of ON–OFF retinotopy). We then used the 2D discrete FFT (2D-FFT) of the ON–OFF receptive field to estimate the predicted preferred orientation preference (Fig. 1b, right). This population analysis included only horizontal cortical penetrations that had at least five recording sites with receptive fields showing clear ON–OFF segregation (SNR of ON–OFF receptive field $>$ 8) and good orientation selectivity measured with moving bars (orientation selectivity $>$ 0.5; goodness of fit for orientation tuning $>$ 0.6). The peaks in the 2D-FFT also had to be distant from the origin, as otherwise the preferred orientation extracted from the 2D-FFT would be ambiguous.

**Orientation pinwheels and direction fractures.** To investigate a possible relationship between ON–OFF dominance and orientation selectivity at pinwheel centres, we selected horizontal recordings in which orientation changed abruptly. To make our sample of orientation discontinuities as homogeneous as possible, we selected only cortical regions that were completely monocular, responded strongly to all stimulus orientations and had responses with high SNR (SNR $>$ 5). We then measured changes in both orientation selectivity and absolute contrast polarity (OFF or ON dominance) as a function of cortical distance from the region with lowest orientation selectivity (Fig. 4e). To investigate a possible relationship between ON–OFF dominance and abrupt changes in direction preference, we selected sections of horizontal cortical penetrations in which orientation preference changed by $<45°$ but direction preference changed abruptly within $\leq 0.2$ mm (receptive field SNR $>$ 5). We then marked the abrupt changes in direction preference as cortical distance 0 and measured changes in direction preference and spatial location of the strongest subregion within the receptive field as a function of cortical distance. To measure the changes in retinotopic position with the maximum accuracy possible, we did not subtract responses to different stimuli and made all the measurements directly from responses to light stimuli.

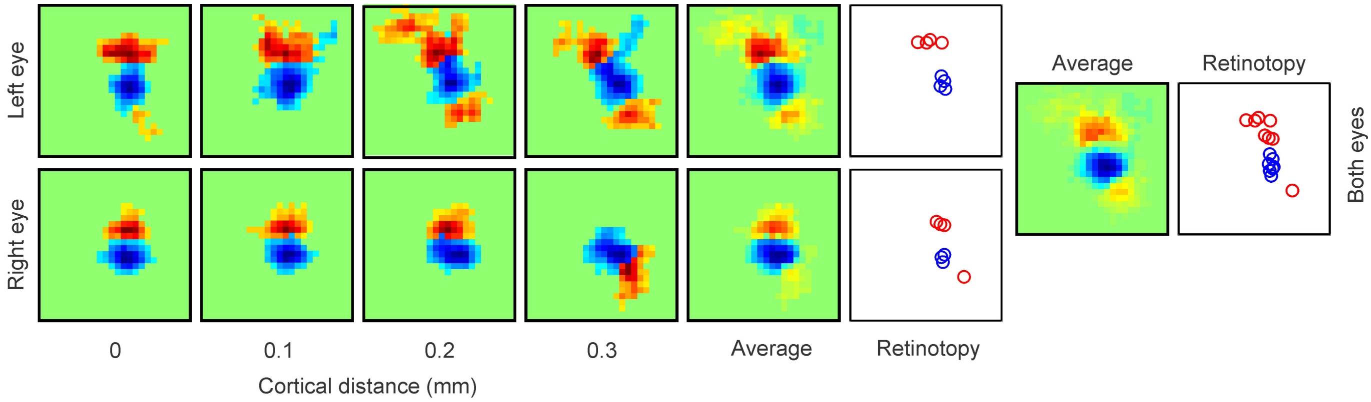

41. Martinez, L. M. *et al.* Receptive field structure varies with layer in the primary visual cortex. *Nature Neurosci.* **8,** 372–379 (2005).
42. Brainard, D. H. The psychophysics toolbox. *Spat. Vis.* **10,** 433–436 (1997).
43. Swindale, N. V., Grinvald, A. & Shmuel, A. The spatial pattern of response magnitude and selectivity for orientation and direction in cat visual cortex. *Cereb. Cortex* **13,** 225–238 (2003).
44. Lashgari, R. *et al.* Response properties of local field potentials and neighboring single neurons in awake primary visual cortex. *J. Neurosci.* **32,** 11396–11413, (2012).
45. DeAngelis, G. C., Ghose, G. M., Ohzawa, I. & Freeman, R. D. Functional micro-organization of primary visual cortex: receptive field analysis of nearby neurons. *J. Neurosci.* **19,** 4046–4064 (1999).



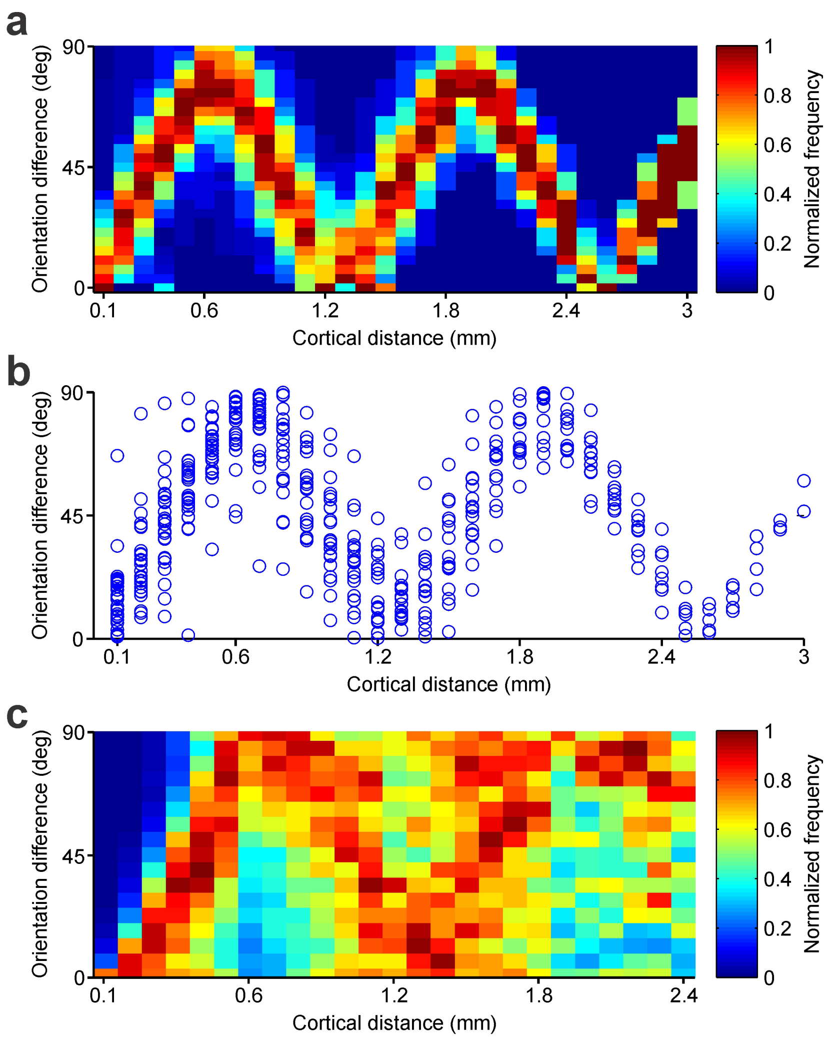

**Extended Data Figure 1 | Measurements of ON–OFF responses and ocular dominance columns. a**, ON and OFF receptive fields were mapped with light (ON) and dark (OFF) sparse noise and calculated from the response to the stimulus onset (grey shaded area). **b**, Horizontal penetrations that ran for more than 1.2 mm through a monocular band were assumed to be nearly parallel to ocular dominance columns (top) and those that alternated between monocular responses for left and right eyes were assumed to be nearly orthogonal to ocular dominance columns (bottom). Receptive fields normalized for ocular dominance. Icons on the left illustrate ocular dominance columns for contralateral (C) and ipsilateral (I) eyes (arrow illustrates horizontal penetration). Each receptive field box has a side of 27°.



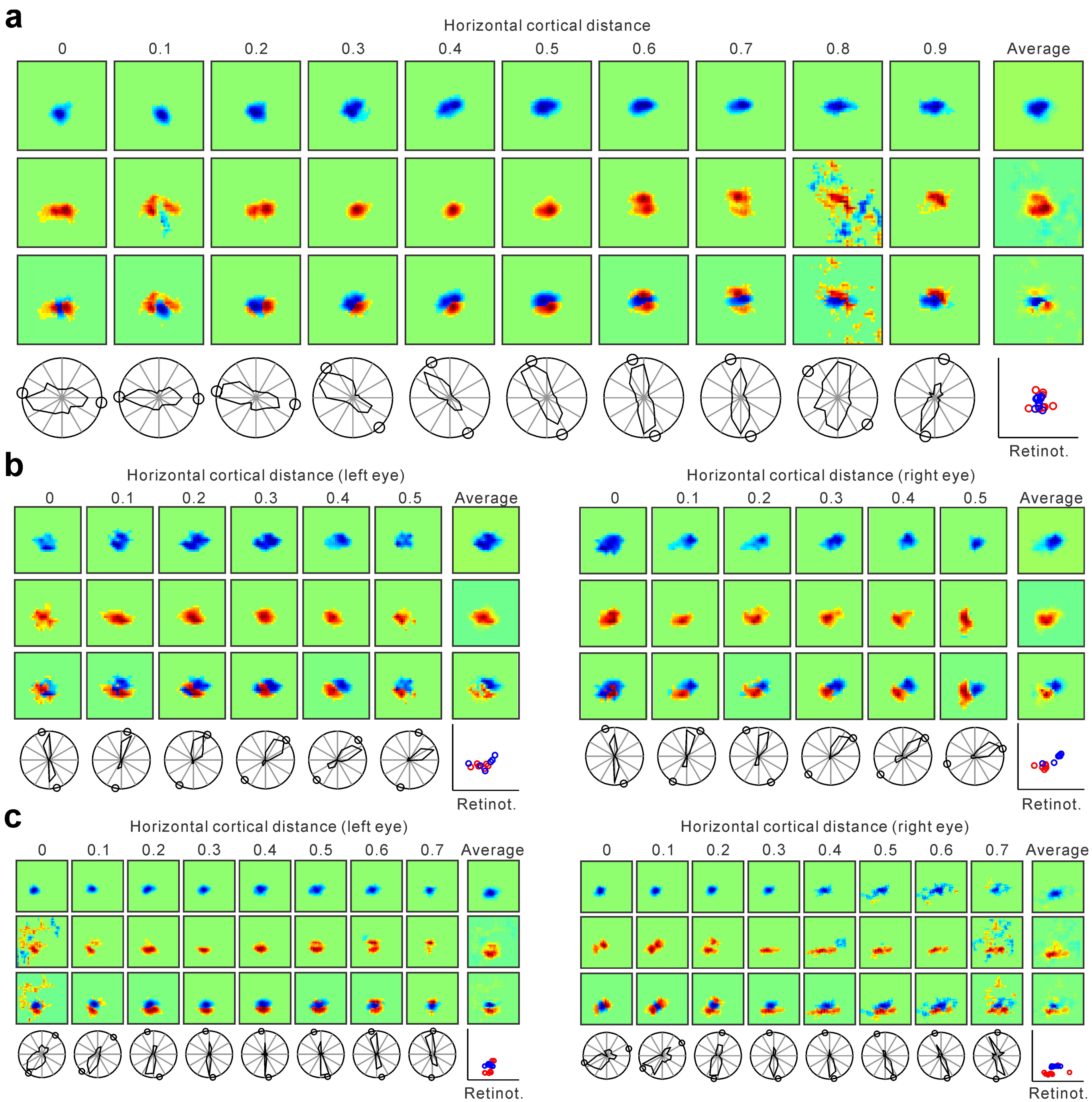

**Extended Data Figure 2 | ON–OFF domains are matched across eyes.** **a**, Integrating the ON–OFF receptive fields over 0.7 mm of horizontal cortical distance reveals ON and OFF receptive field subregions that are segregated in visual space and well matched between eyes. Notice the excellent binocular match of the receptive field subregions measured with light spots (left, two subregions displaced vertically in both eyes), and dark spots (middle left, one central subregion in both eyes). The ON–OFF receptive field difference also shows an excellent binocular match (middle right), so the ON–OFF segregation can still be seen after combining the receptive fields of the two eyes (right). **b**, Integrating the ON–OFF receptive fields over a much longer distance (1.6 mm of cortex, different horizontal penetration) still reveals separate receptive field subregions with excellent binocular match. The 1.6-mm-average receptive fields of the left and right eyes have both two ON subregions that are displaced diagonally and retinotopically matched (left). They also have two OFF subregions that are also displaced diagonally and retinotopically matched between the two eyes (middle left). A hint of the ON subregions can still be seen in the ON–OFF receptive field difference (middle right) and receptive field of both eyes combined (right), even if the receptive fields were averaged over 1.6 mm of cortex. Each square box framing a receptive field has a side of 16.2°.



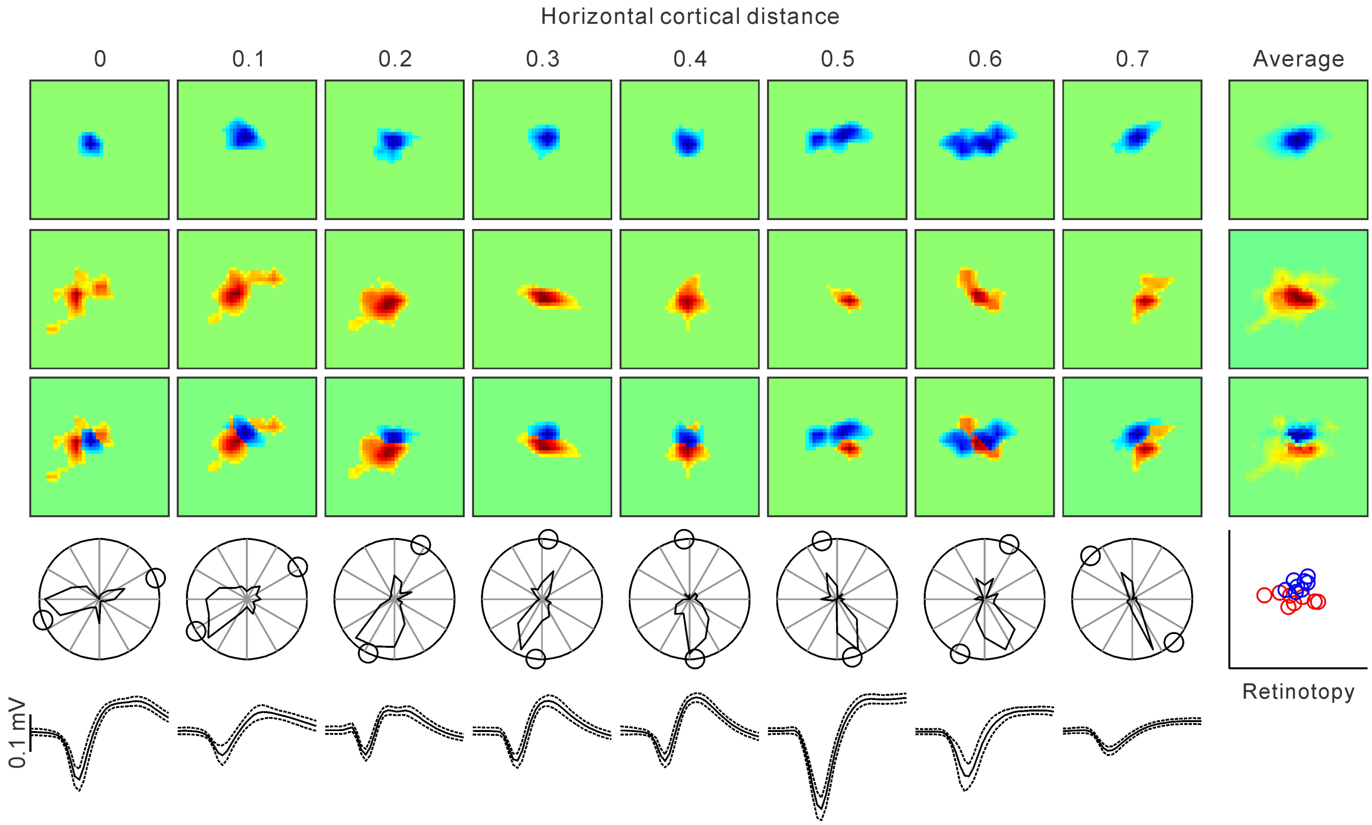

**Extended Data Figure 3 | The OFF pathway might also anchor retinotopy in the primary visual cortex of the macaque.** ON–OFF retinotopy measured along 0.3 mm of horizontal cortical distance in macaque primary visual cortex ($n = 1$ monkey). As in the cat, changes in OFF retinotopy are more restricted than changes in ON retinotopy in the receptive fields of both eyes. Panels labelled 'average' show receptive fields averaged across cortical distance separately for each eye and both eyes. Plots labelled 'retinotopy' show the retinotopy of the receptive field pixel that generated the strongest ON (red) or OFF (blue) response, shown separately for each eye and both eyes. Each square box framing a receptive field has a side of 12°.



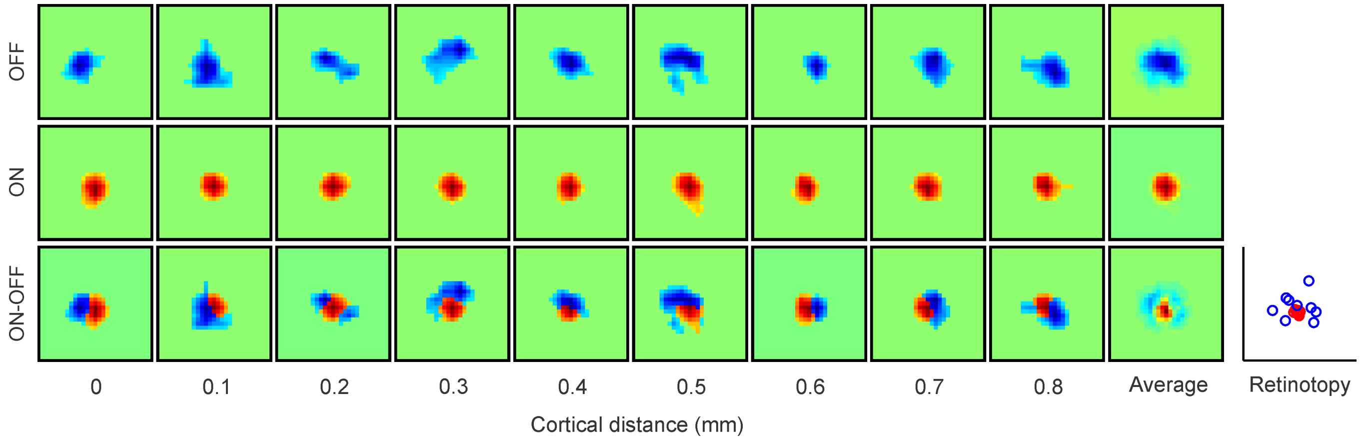

**Extended Data Figure 4 | Periodic changes in orientation preference.** **a**, Colour map showing normalized frequency of orientation difference between paired recordings measured at different cortical distances within a single horizontal penetration (same as Fig. 3k left). **b**, Difference in orientation preference between all possible paired recordings measured within the same horizontal penetration as in **a** ($n = 496$ paired comparisons, $n = 1$ animal). **c**, Same as **a** but for multiple recording sites obtained from multiple penetrations ($n = 20{,}672$ paired comparisons, $n = 36$ animals).




**Extended Data Figure 5 | Additional examples of horizontal recordings showing a correlation between changes in ON–OFF retinotopy and orientation preference. a**, Horizontal recording through 0.9 mm of cortex. From top to bottom, the first three panel rows show series of OFF, ON and ON–OFF receptive fields (left) and receptive fields averaged across horizontal cortical distance (right). The bottom row shows the orientation or direction tuning (left) and the retinotopy (Retinot.) of the strongest response within each receptive field (right; ON, red; OFF, blue). The small circles in the orientation plots illustrate the preferred orientation predicted from the ON–OFF receptive field. **b**, **c**, Horizontal recordings through binocular regions of length 0.5 mm (**b**) and 0.7 mm (**c**). Notice the accurate binocular match in ON–OFF retinotopy between the two eyes and also the striking binocular similarity in orientation preference, direction preference and orientation and direction selectivity. Each receptive field box has a side of 27° (**a**), 23° (**b**) or 23.6° (**c**).




**Extended Data Figure 6 | Example of a horizontal penetration in which we recorded from several single neurons separated from each other by 0.1 mm.** Format is similar to Fig. 4a and Extended Data Fig. 5a. The only difference is that the receptive fields and orientation plots were obtained from single neurons instead of multiunit activity. The last row shows spike waveforms from each single neuron (average and s.d.). Each square box framing a receptive field has a side of 23°.




**Extended Data Figure 7 | Example of a cortical region in which OFF retinotopy rotates around ON retinotopy.** The figure shows a series of receptive fields mapped with dark (OFF) and light stimuli (ON) and the ON–OFF receptive field difference. The last receptive field on the right for each row shows the average of all receptive fields across 0.8 mm of cortical distance. The plot on the right shows the retinotopy of the ON (red) and OFF (blue) receptive fields. Cortical regions where OFF retinotopy rotated around ON retinotopy were more difficult to find than regions where ON retinotopy rotated around OFF retinotopy. To estimate the relative frequency of ON and OFF retinotopy rotations, we measured the distance between the retinotopic centre of mass of single horizontal penetrations for each ON or OFF receptive field (81 penetrations with receptive field measurements from at least five recording sites per penetration). We then calculated a ratio of the average distances, as (ON − OFF)/(ON + OFF), and used a ratio of 0.5 as an arbitrary threshold to classify a penetration as OFF-anchored (ON rotates around OFF) or ON-anchored (OFF rotates around ON). Based on this criterion, there were 3.75 more OFF-anchored than ON-anchored penetrations (15 versus 4 penetrations, respectively; $n = 17$ animals). Each square box framing a receptive field has a side of 19.4°.

doi:10.1038/nature17938

# Continuous evolution of *Bacillus thuringiensis* toxins overcomes insect resistance

Ahmed H. Badran[1,2], Victor M. Guzov[3], Qing Huai[3], Melissa M. Kemp[3], Prashanth Vishwanath[3], Wendy Kain[4], Autumn M. Nance[5], Artem Evdokimov[5]†, Farhad Moshiri[5], Keith H. Turner[5], Ping Wang[4], Thomas Malvar[5] & David R. Liu[1,2]

**The *Bacillus thuringiensis* δ-endotoxins (Bt toxins) are widely used insecticidal proteins in engineered crops that provide agricultural, economic, and environmental benefits. The development of insect resistance to Bt toxins endangers their long-term effectiveness. Here we have developed a phage-assisted continuous evolution selection that rapidly evolves high-affinity protein–protein interactions, and applied this system to evolve variants of the Bt toxin Cry1Ac that bind a cadherin-like receptor from the insect pest *Trichoplusia ni* (TnCAD) that is not natively bound by wild-type Cry1Ac. The resulting evolved Cry1Ac variants bind TnCAD with high affinity (dissociation constant $K_d = 11$–$41$ nM), kill TnCAD-expressing insect cells that are not susceptible to wild-type Cry1Ac, and kill Cry1Ac-resistant *T. ni* insects up to 335-fold more potently than wild-type Cry1Ac. Our findings establish that the evolution of Bt toxins with novel insect cell receptor affinity can overcome insect Bt toxin resistance and confer lethality approaching that of the wild-type Bt toxin against non-resistant insects.**

The expression of insecticidal proteins from *B. thuringiensis* (Bt toxins) in crops has proved to be a valuable strategy for agricultural pest management[1]. Bt-toxin-producing crops have been widely adopted in agriculture with substantial economic and environmental benefits[2], and have increased global agricultural productivity by an estimated US$78 billion from 1996 to 2013 (ref. 3). Unfortunately, Bt toxin resistance has evolved among insect pests and threatens the continued success of this strategy for pest control[4]. While resistance management strategies have been developed, including the use of multiple Bt toxins and preserving susceptible alleles in insect populations, the evolution of insect resistance to Bt toxins remains the most serious current threat to sustaining the gains offered by transgenic crops[4].

Bt toxins interact with protein receptors on the surface of insect midgut cells, leading to pore formation in the cell membrane and cell death[5]. Bt toxin resistance is commonly associated with the mutation, downregulation, or deletion of these receptors[2]. We hypothesized that it might be possible to overcome Bt toxin resistance by evolving novel Bt toxins that bind with high affinity to new gut cell receptor proteins in insects. If successful, such an approach has the potential to alter toxin specificity, improve toxin potency, and bypass receptor-related resistance mechanisms.

Here we use phage-assisted continuous evolution (PACE) to rapidly evolve Bt toxins through more than 500 generations of mutation, selection, and replication to bind a new receptor expressed on the surface of insect midgut cells. PACE-derived Bt toxins bind the new receptor with high affinity and specificity, induce target receptor-dependent lysis of insect cells, and enhance the insecticidal activity against both sensitive and Bt-resistant insect larvae up to 335-fold. Collectively, these results establish an approach to overcoming Bt toxin resistance and provide a new platform for the rapid evolution of other protein-binding biomolecules.

## Development of protein-binding PACE

PACE has mediated the rapid laboratory evolution of diverse protein classes including polymerases, proteases, and genome-editing proteins, yielding variants with highly altered activities and specificities[6–12]. While PACE has not been previously used to evolve protein-binding activity, we speculated that the bacterial two-hybrid system[13] could serve as the basis of a protein-binding PACE selection (Fig. 1a). Target binding results in localization of RNA polymerase upstream of a reporter gene, initiating gene expression. To adapt this system into a protein-binding selection for PACE, we envisioned that protein:target binding could instead activate the expression of the filamentous bacteriophage gene III, which is required for the infectivity of progeny phage[6] (Fig. 1b).

To maximize the sensitivity of the bacterial two-hybrid, we extensively optimized parameters including (1) transcriptional activation and DNA-binding domains, (2) protein expression level, (3) interaction binding affinity, (4) DNA-binding domain multivalency state, (5) reporter gene ribosome-binding site, (6) operator–promoter distance, (7) RNA polymerase–promoter affinity, and (8) DNA-binding domain–bait linker length. While the previously described bacterial two-hybrid system yielded a 17-fold increase in transcriptional activation using a model high-affinity interaction (HA4 monobody binding to the SH2 domain of ABL1 kinase)[14], our optimized system enhanced transcriptional activation >200-fold using the same interaction (Extended Data Figs 1–3). This system consists of the *Escherichia coli* RNA polymerase omega subunit (RpoZ) as the activation domain, the 434 phage cI repressor as the DNA-binding domain, and an optimized $P_{lacZ}$-derived promoter ($P_{lacZ\text{-}opt}$) to drive reporter transcription. Together, these results extend and improve previously described bacterial systems[13] that transduce protein-target binding into gene expression in a manner that can be tuned by the researcher.

[1]Department of Chemistry and Chemical Biology, Harvard University, Cambridge, Massachusetts 02138, USA. [2]Howard Hughes Medical Institute, Harvard University, Cambridge, Massachusetts 02138, USA. [3]Monsanto Company, 245 First Street, Suite 200, Cambridge, Massachusetts 02142, USA. [4]Department of Entomology, Cornell University, Geneva, New York 14456, USA. [5]Monsanto Company, 700 Chesterfield Parkway West, Chesterfield, Missouri 63017, USA. †Present address: HarkerBIO, 700 Ellicott Street, Buffalo, New York 14023, USA.

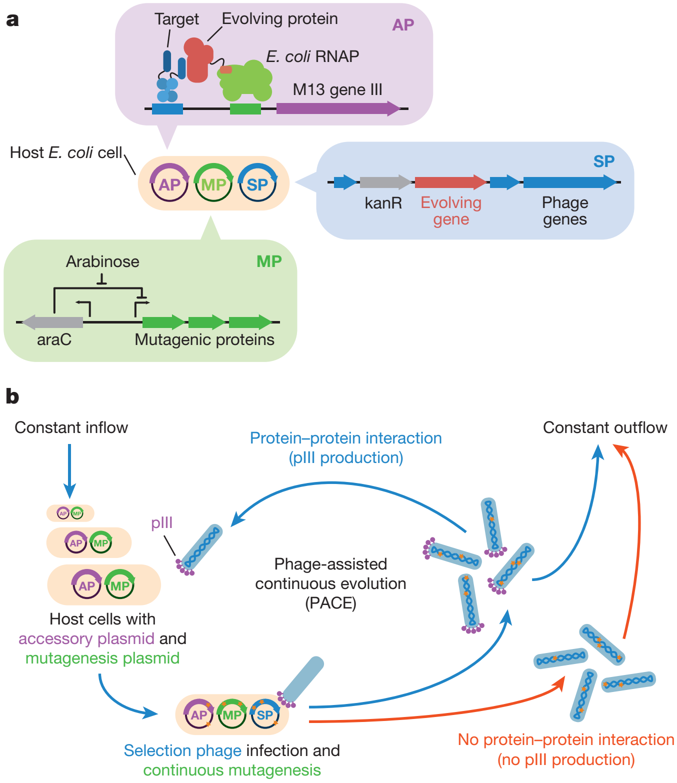



**Figure 1 | Protein-binding PACE. a**, Anatomy of a phage-infected host cell during PACE. The host *E. coli* cell carries two plasmids: the accessory plasmid (AP)[6], which links protein binding to phage propagation and controls selection stringency, and the mutagenesis plasmid (MP)[6,12], which enables arabinose-inducible elevated levels of mutagenesis during PACE. **b**, After infection, each selection phage (SP) that encodes an evolving protein capable of binding to the target protein induces expression of gene III from the accessory plasmid, resulting in the production of pIII, a phage protein required for progeny phage produced by that host cell to infect subsequent host cells. PACE takes place in a fixed-volume vessel (the 'lagoon') that is continuously diluted with fresh host cells. Only those selection phages encoding proteins that bind the target can propagate faster than they are diluted out of the lagoon.

The HA4 monobody binds to the SH2 domain of ABL1 kinase ($K_d = 7$ nM)[14]. The mutant HA4 Y87A monobody binds the ABL1 SH2 domain with 100- to 1,000-fold weaker affinity[14]. Whereas wild-type HA4 monobody fused to RpoZ in the presence of 434cI–SH2 resulted in potent transcriptional activation in our optimized bacterial two-hybrid system, transcriptional activation using $HA4_{Y87A}$ was negligible (Fig. 2a). Similarly, selection phage expressing the *rpoZ–HA4* fusion robustly propagate using host cell strains carrying accessory plasmids expressing the *434cI–SH2* fusion, whereas a selection phage encoding *rpoZ–*$HA4_{Y87A}$ did not support phage propagation. These findings demonstrate that the Y87A mutant HA4 monobody does not support gene III expression or phage propagation (Extended Data Fig. 3).

To validate protein-binding PACE, we challenged the system to evolve a functional SH2-binding monobody starting from the $HA4_{Y87A}$ mutant. To revert the $HA4_{Y87A}$ mutant back to a Tyr87 protein requires three adjacent point mutations (GCG to TAT or TAC). The *rpoZ–*$HA4_{Y87A}$ selection phage was propagated during PACE for 66 h in the absence of selection pressure (that is, allowing evolutionary drift[9]), before engaging selection pressure by changing the host cell strain to one requiring SH2 binding-dependent phage propagation (Fig. 2b). Under selection pressure, control lagoons that previously experienced neither drift nor mutagenesis, or that experienced only mutagenesis, quickly lost their selection phages encoding the evolving monobody population (phage 'wash out'). In contrast, lagoons subjected to both drift and mutagenesis dropped markedly in phage titre for the first 12 h, but recovered over the next 24 h (Fig. 2c). Sequence analysis of eight phage clones surviving 48 h of PACE revealed that all eight evolved either Tyr or Trp at HA4 position 87 (Fig. 2c), either of which restored transcriptional activation (Extended Data Fig. 3). These results demonstrate that protein-binding PACE can rapidly evolve proteins with target affinity, even when multiple mutations are required to gain protein-binding activity.

## Bt toxin target receptor design

Binding of Bt toxins to protein receptors on insect midgut cells is a critical event in the mechanism of insecticidal activity[5,15]. To develop a strategy to overcome Bt toxin resistance, we sought to evolve Cry1Ac, a widely used Bt toxin, to bind TnCAD, an insect cell membrane cadherin-like receptor from cabbage looper (*T. ni*) that is not natively bound by wild-type Cry1Ac (see Supplementary Discussion). *T. ni* has developed Bt resistance in agricultural settings and has been widely studied for insect resistance to Bt toxins[2].

Previous studies of Cry1Ac binding to cadherin-like receptor proteins from Lepidoptera identified multiple putative toxin binding regions (TBRs) in the cadherin[16,17]. The homologous region of the TBR in TnCAD differs from that of cadherin-like proteins from other lepidopteran species at seven amino-acid positions (Extended Data Fig. 4). To create an evolutionary stepping-stone from cadherin-like proteins that bind Cry1Ac to TnCAD, three residues (F1433, S1436, and A1437) from the TBR of four other lepidopteran species[18–21] were introduced into TnCAD, resulting in an artificial receptor fragment designated TnTBR3 (Extended Data Fig. 4). We constructed accessory plasmids expressing various TnTBR3 fragments fused to 434cI and assessed transcriptional activation levels in the presence of various domains of Cry1Ac fused to RpoZ (Extended Data Fig. 4). Only Cry1Ac containing three domains of the active toxin (residues 1–609) showed weak binding activity for TnTBR3 fragment 3 (TnTBR3-F3) (Extended Data Fig. 4). A selection phage carrying the *rpoZ–Cry1Ac* fusion gene replicated ~100-fold in a host strain carrying the TnTBR3-F3 accessory plasmid after propagation overnight, whereas a control selection phage lacking the *rpoZ–Cry1Ac* fusion did not replicate (Extended Data Fig. 4). These observations identified TnTBR3-F3 as a promising evolutionary stepping-stone to serve as a starting target for continuous evolution in PACE.

## Evolution of Cry1Ac to bind TnCAD

We performed 528 h of PACE on Cry1Ac in four segments while varying mutagenesis levels and selection stringency (Fig. 3a). For the first two segments (0–144 h and 144–276 h), the accessory plasmid expressed the TnTBR3-F3 stepping-stone target fused to 434cI. For the final two segments of PACE (276–396 h and 396–528 h), the accessory plasmid expressed the TnCAD-F3 final target fused to 434cI. To enhance mutagenesis, we used the moderate-potency mutagenesis plasmid MP4 (ref. 12) during PACE for binding to TnTBR3-F3 (PACE segments 1 and 2) in an effort to decrease the likelihood of accessing early mutations that could impair essential features of Cry1Ac beyond target receptor binding. During the final two PACE segments for binding to TnCAD-F3 (PACE segments 3 and 4) we used MP6, which induces a greater mutation rate and broader mutational spectrum than MP4 (ref. 12), as phage washout consistently occurred during TnCAD-F3 PACE attempts with MP4, suggesting that higher levels of mutagenesis were required to access rare Cry1Ac mutational combinations that conferred binding to the final TnCAD-F3 target. We increased selection stringency during PACE by increasing lagoon flow rates and reducing the number of TnTBR3-F3 or TnCAD-F3 fragments participating in Cry1Ac variant recognition (Fig. 2a and Extended Data Fig. 3). Phage surviving 528 h of PACE experienced on average 511 generations of mutagenic replication under selection conditions[6].

Sequencing of individual clones at the end of the first PACE segment (144 h; four copies of TnTBR3-F3 per $P_{lacZ\text{-}opt}$ promoter) revealed a strong consensus of two coding mutations in *Cry1Ac*, and one coding

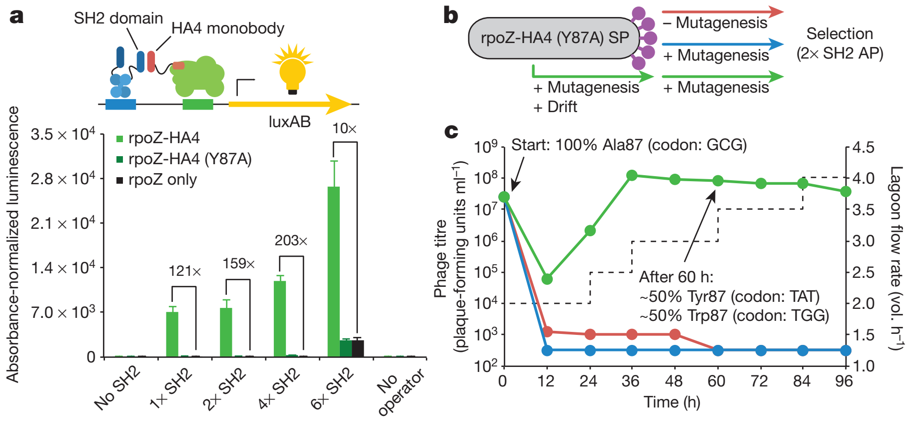



**Figure 2 | Protein-binding PACE selection development and stringency modulation. a**, The relationship between target protein multivalency and transcriptional output measured by luciferase expression. The number of ABL1 SH2 domains available to bind the HA4 monobody was modulated by varying the 434cI DNA-binding domain multivalency state (1×, 2×, 4×, or 6× SH2). 'No operator' indicates a scrambled 434cI operator control accessory plasmid. **b**, During PACE, the inactive monobody mutant $HA4_{Y87A}$ was subjected to no mutagenesis (mutagenesis plasmid not induced), enhanced mutagenesis (mutagenesis plasmid induced with arabinose), or enhanced mutagenesis with genetic drift (mutagenesis plasmid induced with arabinose in addition to an initial period of zero selection stringency), then selected for binding to the ABL1 SH2 target protein. **c**, The combination of drift and enhanced mutagenesis during PACE (green line) resulted in the evolution of Tyr and Trp residues at position 87, either of which restores SH2-binding activity, while no mutagenesis (red line) or enhanced mutagenesis without drift (blue line) resulted in phage washout. Error bars in **a**, s.d. of at least three independent biological replicates.

mutation in *rpoZ* (Extended Data Fig. 5). The three mutations together resulted in 11-fold higher transcriptional activation than that of the wild-type *rpoZ–Cry1Ac* fusion. At the end of the second segment (276 h; two copies of TnTBR3-F3 per $P_{lacZ\text{-}opt}$ promoter), even greater degrees of transcriptional activation were observed, up to 20-fold higher than the level resulting from the starting fusion protein (Fig. 3b and Extended Data Fig. 5). At the end of third segment (396 h; four copies of TnCAD-F3 per $P_{lacZ\text{-}opt}$ promoter), Cry1Ac variants evolved with greatly enhanced apparent affinity for TnCAD-F3 (Fig. 3c and Extended Data Fig. 5). Whereas wild-type Cry1Ac could not detectably activate transcription when challenged to bind TnCAD-F3, single Cry1Ac variants emerging from a total of 384 h of PACE robustly activated transcription up to 210-fold above background in the absence of Cry1Ac variant expression. The end of the fourth segment (528 h; one copy of TnCAD-F3 per $P_{lacZ\text{-}opt}$ promoter) yielded Cry1Ac mutants that could activate transcription when challenged to bind TnCAD-F3 by up to 500-fold (Fig. 3c and Extended Data Fig. 5), consistent with strong binding to the TnCAD-F3 final target.

## Characterization of evolved Cry1Ac variants

DNA sequencing of individual clones surviving 528 h of PACE revealed several consensus genotypes carrying up to 16 mutations per clone out of 22 consensus mutations, most of which localize to domain II, the predicted cadherin-binding domain of Cry1Ac (Extended Data Figs 4 and 5). To illuminate the evolution trajectories en route to TnCAD-F3 binding activity, we analysed all lagoon samples, by high-throughput DNA sequencing using both shorter-read (Illumina) and longer-read (Pacific Biosciences) methods (Extended Data Fig. 6). These efforts identified 25 mutations commonly occurring over the 528 h of PACE (Extended Data Figs 5 and 6). Oligotyping analysis[22] of the long-read data revealed plausible evolutionary trajectories over the entire course of the experiment (Fig. 3d, e). While PACE does not explicitly promote recombination as a mechanism of gene diversification, we observed multiple putative recombination events during the course of Cry1Ac evolution (Fig. 3e). These recombination events, which we presume arose from multiple phage occasionally infecting the same host cell, yielded seminal, highly functional new variants.

On the basis of our mutational analysis, we designed and synthesized consensus Cry1Ac variants containing the most commonly observed mutations (Fig. 4a, b). Purified activated Cry1Ac variants encoding PACE-derived consensus mutations bind strongly ($K_d = 18$–$34$ nM) to a TnCAD fragment containing the TBR (TnCAD-FL; Extended Data Fig. 4) by ForteBio bio-layer interferometry analysis, with evolved Cry1Ac variants C03 and C05 exhibiting the highest binding affinities (Fig. 4a and Supplementary Table 1). In contrast, wild-type Cry1Ac exhibited no significant affinity for TnCAD-FL ($K_d > 1$ mM) under the same conditions. These results together establish the ability of protein-binding PACE to rapidly evolve extensively mutated proteins with high target affinity.

Cry1Ac is proteolytically activated in the insect midgut[2]. The evolved consensus mutants, however, exhibited extensive proteolysis by trypsin under conditions in which the wild-type Cry1Ac was cleanly cleaved into its active form (Fig. 4c). Thermal melting studies confirmed this reduced stability (consensus variants: melting temperature $T_m = {\sim}45$ °C; wild-type Cry1Ac: $T_m = 71$ °C; Supplementary Table 1). Despite this lower stability, trypsin-activated consensus variants robustly killed Sf9 cells expressing TnCAD, whereas wild-type Cry1Ac did not exhibit toxicity (Fig. 4d). Moreover, these evolved consensus Cry1Ac mutants showed insecticidal activity in *T. ni* larvae, although they were less potent than wild-type Cry1Ac (Fig. 4e).

We hypothesized that a subset of the consensus mutations were impairing apparent toxin potency against insect larvae by decreasing Cry1Ac stability and thus promoting degradation in insect gut. We generated Cry1Ac variants containing combinatorial reversions of the identified consensus mutations (Fig. 4b and Supplementary Table 1) and identified mutations D384Y and S404C, two mutations that arose early during PACE against the TnTBR3 stepping-stone target (Figs 3d, e and 5a), as the source of reduced protein stability. Variants lacking these two mutations, but containing the other seven consensus C05 mutations, exhibited greatly improved stability ($T_m = {\sim}60$ °C). Variants lacking D384Y and S404C also exhibited proteolytic resistance similar to that of wild-type Cry1Ac, while retaining high binding affinity to TnCAD-FL ($K_d = 11$–$41$ nM) (Fig. 5a, b and Supplementary Table 1).

We assayed the toxicity of two evolved consensus Cry1Ac variants (C05 and C03) and three stabilized evolved consensus Cry1Ac variants (C05s, C03s, and A01s) lacking D384Y and S404C to cultured Sf9 insect cells expressing an ABCC2 receptor (positive control) or TnCAD. The stabilized evolved Cry1Ac variants retain their ability to bind to the ABCC2 receptor, while acquiring the ability to potently kill Sf9 cells expressing TnCAD, in contrast to the ability of wild-type Cry1Ac to only kill cells expressing the ABCC2 receptor, but not cells expressing TnCAD (Fig. 5c).

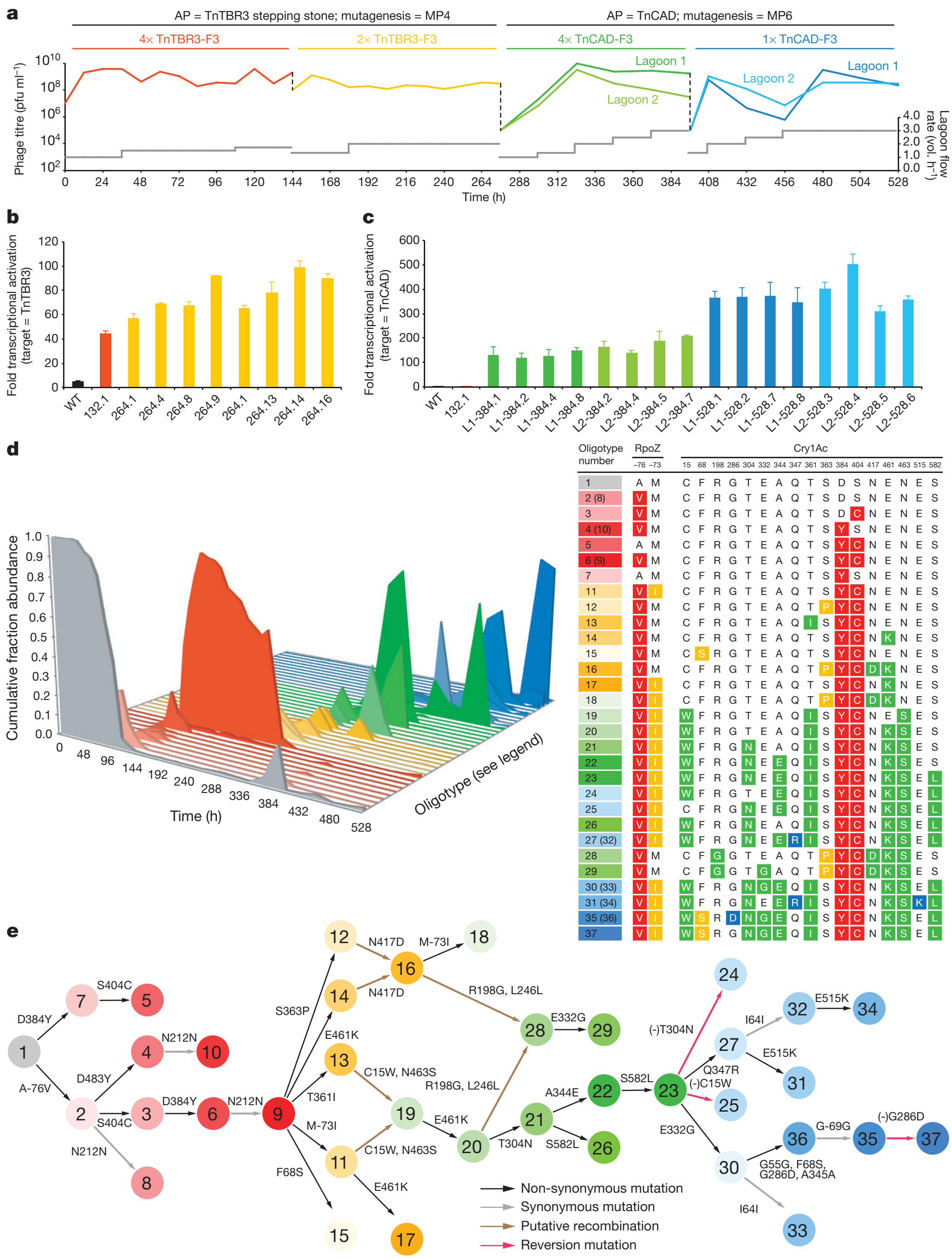


**Figure 3 | Continuous evolution of Cry1Ac variants that bind the *T. ni* cadherin receptor.** **a**, PACE was executed in four segments. The first two segments implemented the designed TnTBR3-F3 'stepping-stone' target under intermediate levels of mutagenesis (MP4). The final two segments implemented the final TnCAD-F3 target under high levels of mutagenesis (MP6). Phage titre (coloured lines) and lagoon flow rate (grey lines) are shown at all sampled time points. The dotted lines indicate transfer of evolving phage to a new lagoon fed by the host cell culture corresponding to the next segment of PACE. **b**, **c**, Transcriptional activation assays using 434cI–TnTBR3-F3 (**b**) or 434cI–TnCAD-F3 (**c**) and individual RpoZ–Cry1Ac variants evolved during PACE, compared with wild-type RpoZ–Cry1Ac (WT). **d**, Oligotyping analysis of lagoon samples during PACE on the basis of high-throughput DNA sequencing data. Oligotypes containing high frequency mutations (≥1%) are represented by different polygons, coloured on the basis of the stage in which they first became abundant in the evolving Cry1Ac gene pool. Mutations in Cry1Ac for each oligotype are shown in the table. Numbers in parentheses indicate the oligotype number assigned to that mutant after a synonymous (silent) mutation. **e**, Plausible evolution trajectories over the entire PACE experiment derived from oligotyping analysis strongly suggests instances of recombination during PACE, and reveals the influence of mutation rate, selection stringency, and target protein on evolutionary outcomes. The colours and numbers in each circle correspond to those in **d**.

## *In vivo* activity of evolved Cry1Ac variants

Finally, we assayed the insecticidal activity of the stabilized evolved Cry1Ac variants against Cry1Ac-sensitive *T. ni* larvae when added to their diet. Consistent with the *in vitro* results, the stabilized evolved Cry1Ac variants exhibited substantially increased toxicity to *T. ni* larvae compared with that of the consensus-evolved Cry1Ac mutants before stabilization (Fig. 5d). Interestingly, the stabilized evolved Cry1Ac variants also exhibited insecticidal potency against susceptible *T. ni* up to fourfold higher than that of wild-type Cry1Ac, suggesting that the evolved affinity of the toxins to a new receptor may augment their insecticidal potency, even against insects susceptible to wild-type Cry1Ac. These results also suggest that the evolution of Bt toxins that recognize novel receptors could expand the range of insects that can be targeted by Bt toxins, consistent with previous *in vitro* studies[23,24] using designed Bt toxin derivatives.

Next we evaluated the insecticidal activity of the stabilized evolved Cry1Ac variants against Cry1Ac-resistant *T. ni* larvae. *T. ni* resistance to Cry1Ac has been genetically mapped to the *ABCC2* transporter gene and downregulation of expression of *APN1* (refs 25, 26), and is known to be independent of alteration of the cadherin-like receptor[27]. In this study, we also confirmed that wild-type Cry1Ac does not bind the TBR in TnCAD (see above), consistent with the previous finding that Cry1Ac does not bind TnCAD in *T. ni* midgut cell membranes[25,27,28]. Indeed, we observed ~1,000-fold lower potency of wild-type Cry1Ac against a Cry1Ac-resistant *T. ni* strain than the potency of wild-type Cry1Ac against susceptible *T. ni*[29] (Fig. 5e). Compared with wild-type Cry1Ac, stabilized evolved Cry1Ac variants C05s, C03s, and A01s showed dramatically improved activity against Cry1Ac-resistant *T. ni*, with median lethal concentration ($LC_{50}$) values up to 335-fold lower than wild-type Cry1Ac (Fig. 5e and Extended Data Table 1).

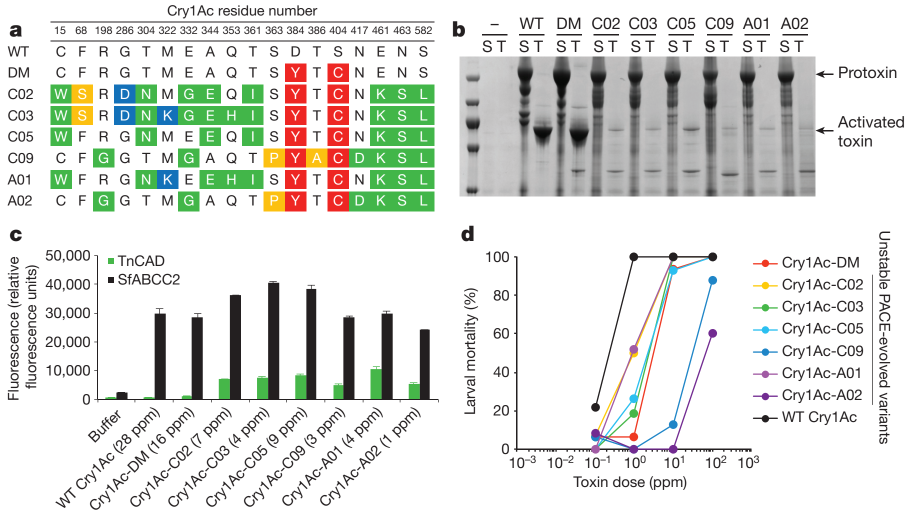



**Figure 4 | Characterization of consensus-evolved Cry1Ac variants.** **a**, Consensus-evolved Cry1Ac mutant sequences, including the D384Y/S404C double mutant (DM) that enabled TnTBR3-F3 recognition during the first segment of PACE. **b**, SDS–polyacrylamide gel electrophoresis (SDS–PAGE) analysis of Cry1Ac variants after trypsin digestion, revealing proteolytic instability of consensus-evolved variants. S, solubilized crude Bt crystals; T, trypsin-treated. **c**, Toxicity assays using Sf9 cells expressing the ABCC2 (black) or TnCAD receptor (green). Cells were incubated with Cry1Ac variants after trypsin digestion at the concentrations of activated toxin shown. Cry1Ac-induced cell permeabilization causes a fluorescent dye to enter cells, resulting in an increase in fluorescence. The evolved Cry1Ac variants, but not wild-type Cry1Ac, induce permeabilization of cells expressing TnCAD. Error bars, s.d. of at least three independent biological replicates. **d**, Insect larvae diet bioassays using wild-type and evolved consensus Cry1Ac variants, showing the loss of evolved Cry1Ac potency in insect larvae arising from impaired stability.

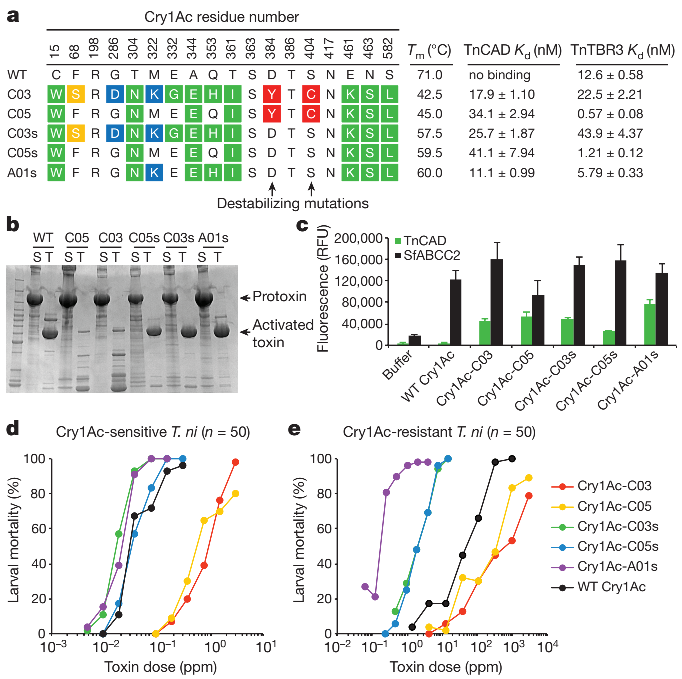



**Figure 5 | Characterization of stabilized evolved Cry1Ac variants reveals potently enhanced activity.** **a**, Sequence, thermal stability, and TnCAD target-binding affinity of unstable and stabilized PACE-evolved consensus mutants. **b**, SDS–PAGE analysis of trypsin digestion reactions showing dramatically enhanced stability upon D384Y and S404C reversion. **c**, Toxicity assays using Sf9 cells overexpressing the ABCC2 (black) or TnCAD receptor (green), demonstrating maintained activity of stabilized variants against both ABCC2 or TnCAD. All variants were used at 10 ppm. Error bars, s.d. of at least three independent biological replicates. **d**, **e**, Highly purified wild-type Cry1Ac, evolved consensus variants, or stabilized evolved variants were added to the diets of Cry1Ac-susceptible (**d**) or Cry1Ac-resistant (**e**) *T. ni* larvae at the indicated doses. Stabilized evolved variants moderately enhance mortality in Cry1Ac-susceptible larvae compared with wild-type Cry1Ac. Stabilized evolved variants greatly outperform wild-type Cry1Ac toxin in killing Bt toxin-resistant *T. ni* larvae.

Importantly, these evolved and stabilized Cry1Ac variants showed similar toxicity in Bt-resistant *T. ni* ($LC_{50} = 0.15$ ppm) as that of wild-type Cry1Ac in susceptible larvae ($LC_{50} = 0.04$ ppm) (Fig. 5e and Extended Data Table 1). Taken together, these results establish that the evolution of novel receptor binding among Bt toxins can overcome Bt toxin resistance in an agricultural pest.

To characterize the species profile of their insecticidal activity, we tested the evolved Cry1Ac variants in diet bioassays against 11 additional agricultural pests: a lepidopteran related to *T. ni* (*Chrysodeixis includes*, soybean looper) that encodes a cadherin-like receptor highly homologous to TnCAD, eight more distantly related lepidopteran pests, and three non-lepidopteran pests (Extended Data Figs 7 and 8). As expected, the stabilized evolved Cry1Ac variants were more potent than wild-type Cry1Ac against *C. includes*, and comparably potent as wild-type Cry1Ac against the other pests assayed (Extended Data Fig. 7). These results further support the mechanism of action of the PACE-evolved Bt toxins as binding to the cadherin receptor in *T. ni* and the closely related cadherin receptor in *C. includes*. Notably, the evolved Bt toxins did not acquire new activity against species lacking a receptor homologous to TnCAD. Taken together, these findings demonstrate that an evolved Bt toxin that binds a novel target can potently kill closely related insect pest species, while maintaining a similar overall insect spectrum as the parental Bt toxin.

## Discussion

Protein-binding PACE rapidly discovered variants of Cry1Ac that bind with high affinity to the novel receptor TnCAD. Perhaps unsurprisingly, we observed a moderate reduction in stability of the evolved variants compared with wild-type Cry1Ac, as stability was not an implicit requirement of the selection. The two mutations that reduced Cry1Ac stability (D384Y and S404C) arose within the first few days of PACE on the stepping-stone target TnTBR3-F3 and were inherited by virtually all subsequent evolved variants (Fig. 3e). It is tempting to speculate that these mutations broadened the substrate scope of Cry1Ac binding to enable downstream protein evolution, at the expense of stability, but were not required once affinity for TnCAD-F3 evolved. Additional affinity measurements of reverted consensus mutations reveal the key roles of E461K, N463S, and S582L, which evolved in quick succession during the third PACE segment (Fig. 3e), consistent with their contri-

bution to TnCAD binding. All three mutations lie on the same face of Cry1Ac (Extended Data Fig. 5c), albeit in different domains, suggestive of potential direct interaction with the cadherin receptor.

Collectively, our findings establish that the laboratory evolution of novel or enhanced Bt toxin–receptor interactions can overcome insect resistance to Bt toxins. This strategy complements existing approaches to limit the incidence of Bt toxin resistance. The 'gene pyramiding' strategy for resistance management[30], for example, requires the availability of multiple effective toxins with different binding sites in target insects. The refuge strategy[4] necessitates that the resistance is a recessive trait and requires compliance by growers. The engineering of Bt toxins to eliminate the reliance on cadherin receptor interaction for toxin oligomerization has been shown to enhance toxicity against resistant strains of several insects, but also reduces the insecticidal potency of the toxins against sensitive insects[31] and may broaden the target specificity of the toxin.

The approach established here enables targeting of a Bt-resistant pest through the evolution of high-affinity Bt toxin variants that bind a specific target insect protein. In principle, this strategy should be applicable to target a variety of insect pests. While the evolution of insect resistance to an evolved Bt toxin is a likely possibility, this work has the potential to provide access to many new Bt toxins that, individually or in combination, may manage resistance and extend the effectiveness of this important approach to pest control. We also envision that this system may be used to explore potential resistance mechanisms by evolving the receptor in the presence of a Bt toxin, analogous to the recent use of PACE to identify protease inhibitor drug resistance mechanisms[10]. Finally, we note that the ability of protein-binding PACE to rapidly evolve novel protein–protein interactions may prove useful in the discovery or improvement of protein therapeutics.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 6 December 2015; accepted 23 March 2016.**
**Published online 27 April 2016.**

1. Prado, J. R. *et al.* Genetically engineered crops: from idea to product. *Annu. Rev. Plant Biol.* **65,** 769–790 (2014).
2. Pardo-López, L., Soberón, M. & Bravo, A. *Bacillus thuringiensis* insecticidal three-domain Cry toxins: mode of action, insect resistance and consequences for crop protection. *FEMS Microbiol. Rev.* **37,** 3–22 (2013).
3. James, C. *Global Status of Commercialized Biotech/GM Crops: 2014.* ISAAA Brief No. 49 (International Service for the Acquisition of Agri-biotech Applications, 2014).
4. Tabashnik, B. E., Brévault, T. & Carrière, Y. Insect resistance to Bt crops: lessons from the first billion acres. *Nature Biotechnol.* **31,** 510–521 (2013).
5. Adang, M. J., Crickmore, N. & Jurat-Fuentes, J. L. Diversity of *Bacillus thuringiensis* crystal toxins and mechanism of action. *Adv. Insect Physiol.* **47,** 39–87 (2014).
6. Esvelt, K. M., Carlson, J. C. & Liu, D. R. A system for the continuous directed evolution of biomolecules. *Nature* **472,** 499–503 (2011).
7. Dickinson, B. C., Leconte, A. M., Allen, B., Esvelt, K. M. & Liu, D. R. Experimental interrogation of the path dependence and stochasticity of protein evolution using phage-assisted continuous evolution. *Proc. Natl Acad. Sci. USA* **110,** 9007–9012 (2013).
8. Leconte, A. M. *et al.* A population-based experimental model for protein evolution: effects of mutation rate and selection stringency on evolutionary outcomes. *Biochemistry* **52,** 1490–1499 (2013).
9. Carlson, J. C., Badran, A. H., Guggiana-Nilo, D. A. & Liu, D. R. Negative selection and stringency modulation in phage-assisted continuous evolution. *Nature Chem. Biol.* **10,** 216–222 (2014).
10. Dickinson, B. C., Packer, M. S., Badran, A. H. & Liu, D. R. A system for the continuous directed evolution of proteases rapidly reveals drug-resistance mutations. *Nature Commun.* **5,** 5352 (2014).
11. Hubbard, B. P. *et al.* Continuous directed evolution of DNA-binding proteins to improve TALEN specificity. *Nature Methods* **12,** 939–942 (2015).
12. Badran, A. H. & Liu, D. R. Development of potent *in vivo* mutagenesis plasmids with broad mutational spectra. *Nature Commun.* **6,** 8425 (2015).
13. Dove, S. L. & Hochschild, A. Conversion of the omega subunit of *Escherichia coli* RNA polymerase into a transcriptional activator or an activation target. *Genes Dev.* **12,** 745–754 (1998).
14. Wojcik, J. *et al.* A potent and highly specific FN3 monobody inhibitor of the Abl SH2 domain. *Nature Struct. Mol. Biol.* **17,** 519–527 (2010).
15. Gómez, I. *et al.* Role of receptor interaction in the mode of action of insecticidal Cry and Cyt toxins produced by *Bacillus thuringiensis*. *Peptides* **28,** 169–173 (2007).
16. Fabrick, J. A. & Tabashnik, B. E. Binding of *Bacillus thuringiensis* toxin Cry1Ac to multiple sites of cadherin in pink bollworm. *Insect Biochem. Mol. Biol.* **37,** 97–106 (2007).
17. Wu, Y. D. Detection and mechanisms of resistance evolved in insects to cry toxins from *Bacillus thuringiensis*. *Adv. Insect Physiol.* **47,** 297–342 (2014).
18. Xie, R. *et al.* Single amino acid mutations in the cadherin receptor from *Heliothis virescens* affect its toxin binding ability to Cry1A toxins. *J. Biol. Chem.* **280,** 8416–8425 (2005).
19. Hua, G., Jurat-Fuentes, J. L. & Adang, M. J. Bt-R1a extracellular cadherin repeat 12 mediates *Bacillus thuringiensis* Cry1Ab binding and cytotoxicity. *J. Biol. Chem.* **279,** 28051–28056 (2004).
20. Nagamatsu, Y., Koike, T., Sasaki, K., Yoshimoto, A. & Furukawa, Y. The cadherin-like protein is essential to specificity determination and cytotoxic action of the *Bacillus thuringiensis* insecticidal CryIAa toxin. *FEBS Lett.* **460,** 385–390 (1999).
21. Peng, D., Xu, X., Ye, W., Yu, Z. & Sun, M. *Helicoverpa armigera* cadherin fragment enhances Cry1Ac insecticidal activity by facilitating toxin-oligomer formation. *Appl. Microbiol. Biotechnol.* **85,** 1033–1040 (2010).
22. Eren, A. M. *et al.* Oligotyping: differentiating between closely related microbial taxa using 16S rRNA gene data. *Methods Ecol. Evol.* **4,** 1111–1119 (2013).
23. Chougule, N. P. *et al.* Retargeting of the *Bacillus thuringiensis* toxin Cyt2Aa against hemipteran insect pests. *Proc. Natl Acad. Sci. USA* **110,** 8465–8470 (2013).
24. Fujii, Y. *et al.* Affinity maturation of Cry1Aa toxin to the *Bombyx mori* cadherin-like receptor by directed evolution. *Mol. Biotechnol.* **54,** 888–899 (2013).
25. Tiewsiri, K. & Wang, P. Differential alteration of two aminopeptidases N associated with resistance to *Bacillus thuringiensis* toxin Cry1Ac in cabbage looper. *Proc. Natl Acad. Sci. USA* **108,** 14037–14042 (2011).
26. Baxter, S. W. *et al.* Parallel evolution of *Bacillus thuringiensis* toxin resistance in Lepidoptera. *Genetics* **189,** 675–679 (2011).
27. Zhang, X., Tiewsiri, K., Kain, W., Huang, L. & Wang, P. Resistance of *Trichoplusia ni* to *Bacillus thuringiensis* toxin Cry1Ac is independent of alteration of the cadherin-like receptor for Cry toxins. *PLoS ONE* **7,** e35991 (2012).
28. Wang, P. *et al.* Mechanism of resistance to *Bacillus thuringiensis* toxin Cry1Ac in a greenhouse population of the cabbage looper, *Trichoplusia ni*. *Appl. Environ. Microbiol.* **73,** 1199–1207 (2007).
29. Song, X., Kain, W., Cassidy, D. & Wang, P. Resistance to *Bacillus thuringiensis* toxin Cry2Ab in *Trichoplusia ni* is conferred by a novel genetic mechanism. *Appl. Environ. Microbiol.* **81,** 5184–5195 (2015).
30. Carrière, Y., Crickmore, N. & Tabashnik, B. E. Optimizing pyramided transgenic Bt crops for sustainable pest management. *Nature Biotechnol.* **33,** 161–168 (2015).
31. Tabashnik, B. E. *et al.* Efficacy of genetically modified Bt toxins against insects with different genetic mechanisms of resistance. *Nature Biotechnol.* **29,** 1128–1131 (2011).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** This work was supported by National Institutes of Health/National Institute of Biomedical Imaging and Bioengineering R01EB022376, DARPA HR0011-11-2-0003, DARPA N66001-12-C-4207, the Howard Hughes Medical Institute, and the US Department of Agriculture National Institute of Food and Agriculture and Agricultural Research Service Biotechnology Risk Assessment Grant Program 2012-33522-19791. A.H.B. was supported by the Harvard Chemical Biology Program and a National Science Foundation Graduate Research Fellowship. We are grateful to J. Carlson, J. Nageotte, D. Rappoli, J.-L. Kouadio, M. Zheng, J. Milligan, M. Huang, Z. Du, X. Zhou, E. Kraft, and J. Wang for their assistance.

**Author Contributions** A.H.B. designed the research, performed experiments, analysed data, and wrote the manuscript. D.R.L. designed and supervised the research and wrote the manuscript. V.M.G. and T.M. designed the initial Cry1Ac/TBR3 pair for evolution in PACE. V.M.G. designed and supervised the research on the evaluation of evolved and stabilized Cry1Ac variants. V.M.G. and P.V. designed stabilized Cry1Ac variants. Q.H. performed protein purification and *in vitro* binding analysis, and analysed data. M.M.K. performed the insect cell-based assays. W.K., P.W., and A.M.N. performed insect diet bioassays using evolved Cry1Ac variants. A.E. and F.M. designed and validated the Cry1Ac-binding TBR3 mutant. K.H.T. analysed high-throughput sequencing data. All of the authors contributed to editing the manuscript.

**Author Information** Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to D.R.L. (drliu@fas.harvard.edu).

## METHODS

No statistical methods were used to predetermine sample size. The *in vivo* experiments were blinded and randomized.

**General methods.** PCR was performed using PfuTurbo Cx Hotstart DNA polymerase (Agilent Technologies), VeraSeq ULtra DNA polymerase (Enzymatics), or Phusion U Hot Start DNA Polymerase (Life Technologies). Water was purified using a MilliQ water purification system (Millipore). Plasmids and selection phages were constructed using USER cloning (New England Biolabs). Genes were either synthesized as bacterial codon-optimized gBlocks Gene Fragments (Integrated DNA Technologies) or amplified by PCR from native sources. *Cry1ac* was amplified by PCR from the *B. thuringiensis* strain Bt_B107284 and cloned into the Bt expression vector pMON101647 using Hot Fusion[32] to generate the expression plasmid pMON133051, which served as a template for amplifying *Cry1ac* fragments for constructing PACE vectors. The toxin-binding region from *T. ni* cadherin (A1133–T1582, AEA29692.10), referred to as TnCAD-FL, was synthesized using 45–60-base oligonucleotides (Integrated DNA Technologies) by overlap extension PCR using KOD Hot Start DNA polymerase (EMD Millipore). The synthetic wild-type TnCAD-FL template was used to generate the TnTBR3-FL fragment via site-directed mutagenesis using the QuikChange II kit according to the manufacturer's instructions (Agilent Technologies). DNA vector amplification was performed using NEB Turbo or DH5$\alpha$ cells (New England Biolabs).

**Electrocompetent strain preparation.** The previously described strains S1030 (ref. 9) or S2060 (ref. 11) were used in all luciferase and plaque assays, as well as in PACE experiments. The glycerol stock of either strain was used to seed a 2-ml overnight culture using 2xYT media (United States Biological) supplemented with 10 $\mu$g ml$^{-1}$ tetracycline (Sigma Aldrich), 50 $\mu$g ml$^{-1}$ streptomycin (Sigma Aldrich), 10 $\mu$g ml$^{-1}$ fluconazole (TCI America), and 10 $\mu$g ml$^{-1}$ amphotericin B (TCI America) in a 37 °C shaker at 230 r.p.m. The saturated culture was diluted 1,000-fold in 50 ml of the same supplemented media and grown under identical conditions until it reached mid-log-phase (absorbance at 600 nm ($A_{600\,\mathrm{nm}}$) = 0.5–0.8). Once the appropriate $A_{600\,\mathrm{nm}}$ was reached, the cells were pelleted in a 50-ml conical tube (VWR) centrifuged at 10,000*g* for 5 min at 4 °C. The supernatant was immediately decanted and the interior of the tube was wiped with a few Kimwipes (Kimberly-Clark) to remove residual media and salts. The cells were resuspended in 25 ml of pre-chilled, sterile filtered 10% glycerol in MilliQ purified water using a pipette to quickly break up the pellet. The cells were centrifuged and washed an additional three times. After the last centrifugation step, the interior of the tube was wiped with a few Kimwipes to remove residual glycerol solution. The pellet was resuspended in as little volume as possible, typically ~150 $\mu$l, and split into 10 $\mu$l aliquots for storage. Cells were flash frozen using a liquid $N_2$ bath, then quickly transferred to −80 °C for extended storage. Electrocompetent S1030 or S2060 cells produced by this method typically yielded $10^7$–$10^8$ colonies per microgram plasmid DNA and enable the simultaneous electroporation of up to three plasmids carrying orthogonal origins of replication and antibiotic resistance cassettes to yield transformants containing all plasmids.

**General USER cloning.** All PACE-related plasmids and phage materials were constructed via USER cloning[33] (see Extended Data Table 2). Briefly, primers were designed to include a single internal deoxyuracil base 15–20 bases from the 5′ end of the primer, specifying this region as the 'USER junction'. Criteria for design of the USER junction were: it should contain minimal secondary structure, have 45 °C < $T_m$ < 70 °C, and begin with a deoxyadenosine and end with a deoxythymine (to be replaced by deoxyuridine). The USER junction specifies the homology required for correct assembly. We note that PfuTurbo Cx Hotstart DNA polymerase (Agilent Technologies), VeraSeq ULtra DNA polymerase (Enzymatics), or Phusion U Hot Start DNA Polymerase (Life Technologies) are able to use primers carrying deoxyuracil bases, whereas some other polymerases undergo a phenomenon known as PCR poisoning and do not extend the primer.

All PCR products were purified using a MinElute PCR Purification Kit (Qiagen) to 10 $\mu$l final volume and quantified using a NanoDrop 1000 Spectrophotometer (Thermo Scientific). For assembly, PCR products carrying complementary USER junctions were mixed in an equimolar ratio (up to 1 pmol each) in a 10 $\mu$l reaction containing 15 units DpnI (New England Biolabs), 0.75 units USER (Uracil-Specific Excision Reagent) enzyme (Endonuclease VIII and Uracil-DNA Glycosylase, NEB), 50 mM potassium acetate, 20 mM Tris-acetate, 10 mM magnesium acetate, 100 $\mu$g ml$^{-1}$ BSA at pH 7.9 (1× CutSmart Buffer, New England Biolabs). The reactions were incubated at 37 °C for 45 min, followed by heating to 80 °C and slow cooling to 22 °C at 0.1 °C s$^{-1}$ in a temperature-controlled block. The hybridized constructs were directly used for heat-shock transformation of chemically competent NEB Turbo *E. coli* cells according to the manufacturer's instructions. Transformants were selected on 1.8% agar-2xYT plates supplemented with the appropriate antibiotic(s).

For selection phage cloning, the hybridized constructs were purified using EconoSpin purification columns (Epoch Life Sciences), eluted using 25 $\mu$l 10% glycerol, and transformed into electrocompetent S2060 cells carrying the phage-responsive accessory plasmid pJC175e, which produces functional pIII in response to phage infection (this strain is henceforth referred to as S2208). After recovery for 3–4 h at 37 °C using 2xYT (United States Biological) media, the culture was centrifuged and the supernatant was purified using a 0.22 $\mu$m PVDF Ultrafree centrifugal filter (Millipore). The supernatant was diluted serially in 100-fold increments and used in plaque assays using log-phase S2208 cells. After overnight at 37 °C, single plaques were picked into 2xYT media and grown for 12–18 h in a 37 °C shaker at 230 r.p.m. The supernatant was purified again to yield clonal phage stocks. In all cases, cloned plasmids and phages were prepared using the TempliPhi 500 Amplification Kit (GE Life Sciences) according to the manufacturers protocol and verified by Sanger sequencing.

**Plaque assays.** S1030 (ref. 9) or S2060 (ref. 11) cells transformed with the accessory plasmid of interest were grown in 2xYT (United States Biological) liquid media supplemented with the appropriate antibiotics to an $A_{600\,\mathrm{nm}}$ of 0.6–0.9. The phage supernatant was diluted serially in three, 100-fold increments to yield four total samples (undiluted, $10^2$-, $10^4$-, and $10^6$-fold diluted) to be used for infections. For each sample, 150 $\mu$l of cells were added to 10 $\mu$l of phage that had been filtered using a 0.22 $\mu$m PVDF Ultrafree centrifugal filter (Millipore). Within 1–2 min of infection, 1 ml of warm (~55 °C) top agar (7 g l$^{-1}$ bacteriological agar in 2xYT) was added to the phage/cell mixture, mixed by pipetting up and down once, and plated onto quartered plates that had been previously poured with 2 ml of bottom agar (18 g l$^{-1}$ bacteriological agar in 2xYT) in each quadrant. The plates were then grown overnight at 37 °C before plaques could be observed.

**PACE.** Host cell cultures, lagoons, media, and the PACE apparatus were as previously described[9]. Recombined selection phage harbouring gene III (rSP) will poison a PACE experiment by outcompeting the evolving selection phage. We have noted that the likelihood of rSP occurrence in a selection phage stock increases with extended standing culture growth during the initial selection phage stock preparation. To reduce the likelihood of rSP formation, all selection phages are repurified before any continuous evolution experiments. Briefly, selection phages were plaqued on S2208 cells. A single plaque was picked into 2 ml 2xYT (United States Biological) supplemented with the appropriate antibiotics and grown until the culture reached mid-log-phase ($A_{600\,\mathrm{nm}}$ = 0.5–0.8). The culture was centrifuged using a table-top centrifuge for 2 min at 10,000*g*, followed by supernatant filtration using a 0.22 $\mu$m PVDF Ultrafree centrifugal filter (Millipore). This short growth period routinely yields titres of $10^6$–$10^8$ plaque-forming units per millilitre and was found to minimize the occurrence of rSP during PACE experiments.

To prepare the PACE strain, the accessory plasmid and MP were co-transformed into electrocompetent S1030 cells (see above) and recovered using Davis rich media[9] (DRM) to ensure MP repression. Transformations were plated on 1.8% agar-2xYT containing 50 $\mu$g ml$^{-1}$ carbenicillin, 40 $\mu$g ml$^{-1}$ chloramphenicol, 10 $\mu$g ml$^{-1}$ fluconazole, 10 $\mu$g ml$^{-1}$ amphotericin B, 100 mM glucose (United States Biological) and grown for 12–18 h in a 37 °C incubator. After overnight growth, four single colonies were picked and resuspended in DRM, then serially diluted and plated on 1.8% agar-2xYT containing 50 $\mu$g ml$^{-1}$ carbenicillin, 40 $\mu$g ml$^{-1}$ chloramphenicol, 10 $\mu$g ml$^{-1}$ fluconazole, 10 $\mu$g ml$^{-1}$ amphotericin B, and either 100 mM glucose or 100 mM arabinose (Gold Biotechnology) and grown for 12–18 h in a 37 °C incubator. Concomitant with this plating step, the dilution series was used to inoculate liquid cultures in DRM supplemented with 50 $\mu$g ml$^{-1}$ carbenicillin, 40 $\mu$g ml$^{-1}$ chloramphenicol, 10 $\mu$g ml$^{-1}$ tetracycline, 50 $\mu$g ml$^{-1}$ streptomycin, 10 $\mu$g ml$^{-1}$ fluconazole, 10 $\mu$g ml$^{-1}$ amphotericin B and grown for 12–18 h in a 37 °C shaker at 230 r.p.m. After confirmation of arabinose sensitivity using the plate assay, cultures of the serially diluted colonies still in log-phase growth were used to seed a 25-ml starter culture for the PACE chemostat.

Once the starter culture had reached log-phase density ($A_{600\,\mathrm{nm}}$ = 0.5–0.8), the 25-ml culture was added directly to 175 ml of fresh DRM in the chemostat. The chemostat culture was maintained at 200 ml and grown at a dilution rate of 1.5–1.6 volumes per hour as previously described[9]. Lagoons flowing from the chemostats were maintained at 40 ml, and diluted as described for each experiment. Lagoons were supplemented with 25 mM arabinose to induce the MP for 8–16 h before infection with packaged selection phage. Samples were taken at the indicated time points, centrifuged at 10,000*g* for 2 min, then filtered with a 0.2 $\mu$m filter and stored overnight at 4 °C. Phage aliquots were titred by plaque assay on S2208 cells (total phage titre) and S1030 or S2060 cells (rSP titre) for all time points.

**Mutagenesis during PACE.** The basal mutation rate of replicating filamentous phage in *E. coli* ($7.2 \times 10^{-7}$ substitutions per base pair per generation) is sufficient to generate all possible single but not double mutants of a given gene in a 40-ml lagoon after one generation of phage replication. For the 2,139-base-pair *rpoZ–Cry1Ac* target, a basal mutation rate of $7.2 \times 10^{-7}$ substitutions per base pair per generation applied to $2 \times 10^{10}$ copies of the gene (a single generation)

in a 40-ml lagoon yields ~$3.1 \times 10^7$ base substitutions, easily enough to cover all 6,417 single point mutants but not all double mutants. Arabinose induction of our first-generation mutagenesis plasmid, MP1, increased the phage mutation rate by ~100-fold, resulting in $7.2 \times 10^{-5}$ substitutions per base pair per generation, yielding ~$3.1 \times 10^9$ substitutions spread over $2 \times 10^{10}$ copies of the gene after a single generation. This enhanced mutation rate is sufficient to cover all possible single mutants ($6.4 \times 10^3$ possibilities) and double mutants ($4.1 \times 10^7$ possibilities), but no triple mutants ($2.6 \times 10^{11}$ possibilities) after a single phage generation. Our recent efforts to enhance mutagenesis in PACE yielded the improved MP6 system[12], which increases the phage mutation rate by an additional 100-fold compared with MP1, resulting in $7.2 \times 10^{-3}$ substitutions per base pair per generation, yielding ~$3.1 \times 10^{11}$ substitutions spread over $2 \times 10^{10}$ copies of the gene after a single generation. This elevated mutation rate is sufficient to cover all possible single mutants ($6.4 \times 10^3$ possibilities), double mutants ($4.1 \times 10^7$ possibilities), and many triple mutants ($2.6 \times 10^{11}$ possibilities) after a single phage generation.

**Luciferase assays.** Complementary plasmids were co-transformed with an accessory plasmid of interest into electrocompetent S1030 (ref. 9) or S2060 (ref. 11) cells and plated onto 1.8% agar-2xYT plates with 50 μg ml$^{-1}$ carbenicillin and 100 μg ml$^{-1}$ spectinomycin. After overnight growth at 37 °C, single colonies were each picked into 2 ml DRM supplemented with 50 μg ml$^{-1}$ carbenicillin, 100 μg ml$^{-1}$ spectinomycin, 10 μg ml$^{-1}$ tetracycline, 50 μg ml$^{-1}$ streptomycin, 10 μg ml$^{-1}$ fluconazole, 10 μg ml$^{-1}$ amphotericin B and grown for 12–18 h in a 37 °C shaker at 230 r.p.m. After overnight growth, cultures were diluted 1,000-fold in a 96-well deep well plate containing 500 μl DRM with 50 μg ml$^{-1}$ carbenicillin, 100 μg ml$^{-1}$ spectinomycin, and the indicated arabinose, isopropyl-β-D-thiogalactoside (IPTG), or anhydrotetracycline (ATc) concentration to induce protein expression from either the accessory plasmid or complementary plasmid. Constitutive accessory plasmids and complementary plasmids were used where no inducer concentration is given. After growth with shaking at 37 °C for 4–5 h, 150 μl of each culture was transferred to a 96-well black wall, clear bottom plate (Costar), and the $A_{600\,\mathrm{nm}}$ and luminescence for each well was measured on an Infinite M1000 Pro microplate reader (Tecan). The $A_{600\,\mathrm{nm}}$ of a well containing only media was subtracted from all sample wells to obtain a corrected $A_{600\,\mathrm{nm}}$ value for each well. The raw luminescence value for each well was then divided by that well's corrected $A_{600\,\mathrm{nm}}$ value to obtain the luminescence value normalized to cell density. Each variant was assayed in at least biological triplicate, and the error bars shown reflect the standard deviations of the independent measurements.

**High-throughput sequencing and oligotype analysis.** Raw reads have been deposited in the NCBI Sequence Read Archive under accession number PRJNA293870, and all custom scripts used in analysis are available at http://github.com/MonsantoCo/BadranEtAl2015. Illumina reads obtained from each time point were mapped to the SP055-rpoZ-cMyc-Cry1Ac1-d123 reference sequence using bowtie version 2.1.0 (ref. 34), and the resulting SAM files were combined into a single BAM file using samtools version 0.1.19 (ref. 35). This BAM file was used as input to freebayes version 0.9.21-12-g92eb53a[36] to call single nucleotide polymorphisms, using the command 'freebayes–use-best-n-alleles 1–pooled-continuous–use-reference-allele–theta 500000000–min-alternate-fraction 0.01–ploidy 1–region SP055-rpoZ-cMyc-Cry1Ac1-d123:2833-4971'. The analysis is encapsulated in the custom script 'ill.callsnps.sh.' PacBio polymerase reads were demultiplexed with RS_Resequencing_Barcode.1 workflow provided by PacBio. Polymerase reads with quality score lower than 0.80 (defined by the PacBio scoring algorithm) or shorter than 50 base pairs were filtered. High-quality reads were processed into subreads after sequencing primers and adaptors were removed. Circular consensus reads (or reads-of-inserts) were obtained by calling consensus of subreads generated from the same polymerase reads. These circular consensus reads were mapped to the SP055-rpoZ-cMyc-Cry1Ac1-d123 reference sequence using BLASR version 1.3.1.142244 (ref. 37), and the alignment was exported as an aligned FASTA sequence using the custom script 'SAMtoAFA.py.' The aligned FASTA was used as input to the oligotyping platform[22], manually specifying entropy components as the positions at which the Illumina data defined informative single nucleotide polymorphisms. Only oligotypes that occur at >1% in at least one sample were retained. This methodology resulted in informative changes at 25 of the 27 specified components. Oligotypes with gaps at the specified components, probably because of indels in the PacBio sequencing or alignment, were reassigned to other oligotypes with nucleotides in those positions only when it could be done unambiguously, and discarded otherwise, resulting in a total fraction abundance <1 in Fig. 5d. The resulting oligotype-percent abundance matrix was read into R and analysed using the custom script 'PedigreeAndMullerPlot.R.' The pedigree was refined manually, assuming that single-mutant derivatives of previous oligotypes were due to *de novo* mutation, while double, triple, or greater mutations that can be explained by recombination of previously observed oligotypes were due to recombination, since these last types of mutation were highly unlikely to arise by multiple point mutation after the start of the PACE experiment.

**High-throughput primary Bt toxin preparation and analysis.** Wild-type *Cry1Ac* was cloned into the Bt expression vector pMON262346 using BspQ1 endonuclease restriction sites. Consensus PACE-evolved *Cry1Ac* variants were synthesized (Gen9) and cloned into the Bt expression vector pMON262346 using Hot Fusion[32]. Reversion mutants of consensus Cry1Ac PACE variants were generated via PCR with Phusion High-Fidelity DNA polymerase (New England Biolabs) and mutant primers followed by Hot Fusion into the Bt expression vector pMON262346. The resulting plasmids were transformed into the protease-deficient Bt strain EG10650 (ref. 38) for protein expression. Cells were grown from single colonies in 96-well plates (Thermo Scientific, AB-0932) overnight in 400 μl Brain Heart Infusion Glycerol (BHIG) media (VWR) supplemented with 5 μg ml$^{-1}$ chloramphenicol. Overnight cultures were used to prepare glycerol stocks (15% glycerol final concentration) and stored at −80 °C for future protein expression. After overnight growth, 10 μl of each culture was used to inoculate 1 ml of complete *C2* medium[39] containing 5 μg ml$^{-1}$ chloramphenicol in 96-well plates. The plates were incubated at 26 °C with vigorous shaking at 550 r.p.m. in a Multitron shaking incubator (Infors HT) for 72 h. The cells were harvested by centrifugation at 3,200*g* for 15 min at 4 °C. The supernatant was decanted and a single 3.5 mm glass bead was added to each well of the plate. The pellet was then resuspended in 1 ml of TX wash buffer composed of 10 mM Tris-HCl, pH 7.5, 0.005% Triton X-100 supplemented with 25 units per millilitre Benzonase (EMD Millipore), and 2 mM $MgCl_2$, incubated at room temperature (21 °C) for 30–60 min (with vigorous vortexing every 10 min), then centrifuged at 3,200*g* for 15 min at 4 °C. The resulting pellet was resuspended and centrifuged under identical conditions two additional times.

The washed spore/crystal pellet from each 1-ml culture was solubilized in the 96-well plate using 300 μl of solubilization buffer composed of 50 mM CAPS, pH 11, and 10 mM DTT, then incubated while shaking at room temperature (21 °C) for 1 h. The insoluble debris was pelleted by centrifugation at 3,200*g* for 15 min at 4 °C, and 200 μl of the supernatant were transferred to a sterile U-bottom 96-well plate. To each well, 10 μl of 0.2 mg ml$^{-1}$ trypsin in 1 M Tris-HCl, pH 7.5 was added. The mixture was incubated at 37 °C for 2 h while shaking at 150 r.p.m., followed by quenching using 2 μl 0.1 M PMSF. The solution was filtered using a Millipore multiscreen plate with a 0.22 μm membrane. Protein stability was assessed by SDS–PAGE and quantified using spot densitometry. Proteins purified using this protocol were tested in downstream insect cell assays.

**Secondary Bt toxin purification and analysis.** Bt glycerol stocks described above were used for large-scale protein expression and purification. A 2-ml starter culture of BHIG medium supplemented with 5 μg ml$^{-1}$ chloramphenicol was inoculated from the glycerol stocks and grown overnight at 280 r.p.m. in a 28 °C shaker. The following day, the saturated culture was transferred into 500 ml complete C2 medium containing 5 μg ml$^{-1}$ chloramphenicol in a 2 l baffled flask and grown for an additional 72 h at 26 °C while shaking at 280 r.p.m. Sporulation and crystal formation in the culture was verified by optical microscopy of a 2-μl aliquot of the saturated Bt culture. Upon confirmation of crystals, the partly lysed sporulated cells were harvested by centrifugation at 10,000*g* for 12 min at 4 °C. The pellet was then resuspended in 100 ml TX wash buffer composed of 10 mM Tris-HCl, pH 7.5, and 0.005% Triton X-100 supplemented with 0.1 mM PMSF, 25 units per millilitre Benzonase (Sigma-Aldrich), and 2 mM $MgCl_2$, incubated at room temperature (21 °C) for 30–60 min (with vigorous vortexing every 10 min), then centrifuged at 3,200*g* for 15 min at 4 °C. The resulting pellet was resuspended and centrifuged under identical conditions two additional times.

The washed spore/crystal pellet was solubilized in 120 ml 50 mM CAPS, pH 11, 10 mM DTT at room temperature for 1 h while shaking at 130 r.p.m. The solubilized protein was separated from the insoluble debris by centrifugation at 35,000*g* for 20 min at 4 °C. The supernatant was transferred to a fresh flask, and then supplemented with 10 ml 0.2 mg ml$^{-1}$ trypsin in 1 M Tris-HCl at pH 7.5. The mixture was incubated at 30 °C for 2–6 h with shaking at 150 r.p.m. and trypsinization was monitored by SDS–PAGE. Once the trypsin digestion reaction was complete, the mixture was centrifuged at 3,200*g* for 15 min at 4 °C. The clear supernatant was removed and mixed with PMSF to 1 mM final concentration. The sample was loaded on a 5–10 ml Q-Sepharose (GE Healthcare) anion exchange column at a flow-rate of 4 ml min$^{-1}$ and the trypsin resistant core of the toxin was eluted in 25 mM sodium carbonate, pH 9 supplemented with 200–400 mM NaCl. Fractions containing the toxin tryptic cores were pooled, concentrated (Millipore Amicon Ultra-15 centrifugal filter Units, Fisher), and loaded on a Hiload Superdex 200 gel filtration column using an ÄKTA chromatography system (GE Healthcare). The column was pre-equilibrated and run with 25 mM sodium carbonate at pH 10.5 supplemented with 1 mM β-mercaptoethanol. Only the monomer peak of the toxin fractions was collected in each case and concentrated to 1–3 mg ml$^{-1}$. The final protein concentration was quantified by spot densitometry. The quality of the trypsinized toxin was assessed using the peptide mass fingerprinting method that was based on in-gel digestion of proteins by trypsin and mass spectrometry analysis of the resulted peptides.

***T. ni* receptor fragment expression and purification.** Custom expression vectors pMON251427 and IS0008 (the same as pMON251427 but with wild-type TnCAD) were used to express 6xHis-TnTBR3-FL and 6xHis-TnCAD-FL fragments in *E. coli*. Both vectors contain an amino (N)-terminal MBP-TVMV protease cleavage site tag[40] and a carboxy (C)-terminal 6× histidine tag flanking the receptor fragment of interest, with the open reading frame driven by the T7 promoter. Expression vectors were transformed into commercial BL21 (λDE3) competent cells (Life Technologies) that had been previously transformed with TVMV protease expression vector (pMON101695; encodes constitutive TVMV protease from a pACYC184 (New England Biolabs) backbone). A single colony was inoculated in 2 ml of Luria-Bertani (LB) media supplemented with 50 μg ml$^{-1}$ kanamycin and 25 μg ml$^{-1}$ chloramphenicol, and grown at 37 °C for 4 h to generate a starter culture, which was used to prepare glycerol stocks and stored at −80 °C for the future protein expression. A second starter culture was inoculated using the BL21 (λDE3) strain glycerol stocks in 2 ml of LB media supplemented with 50 μg ml$^{-1}$ kanamycin and 25 μg ml$^{-1}$ chloramphenicol and grown in a 25 °C shaker (280 r.p.m.) for 15 h. The culture was transferred into 500 ml of Terrific Broth medium (24 g l$^{-1}$ yeast extract, 12 g l$^{-1}$ tryptone, and 5 g l$^{-1}$ glucose) supplemented with 50 μg ml$^{-1}$ kanamycin and 25 μg ml$^{-1}$ chloramphenicol, and grown at 37 °C for 4 h at 280 r.p.m., then transferred to 15 °C and grown for an additional 48 h after supplementation with IPTG to a final concentration of 0.1 mM.

The cells were harvested by centrifugation at 10,000*g* for 12 min at 4 °C. The bacterial cell pellet was resuspended in affinity buffer A (25 mM Tris-HCl at pH 8.0, 0.5 M NaCl, 15 mM imidazole, and 0.2 mM $CaCl_2$) containing 125 units per millilitre of Benzonase (EMD Millipore), 10,000 units per millilitre of chicken egg white lysozyme (Sigma Aldrich) and 1× BugBuster (Novagen). The cell slurry was incubated at room temperature for 15 min, followed by sonication using a Cell Disruptor W-0375 (Heat Systems-Ultrasonics) at 45% Duty Cycle (output number 5) for 30 s with 60 s rests for a total of three cycles. The cell lysate was centrifuged at 35,000*g* for 20 min at 4 °C. The supernatant was loaded onto a 5-ml Ni-NTA column that had been pre-equilibrated using affinity buffer A. After extensive washing with affinity buffer A, the receptor fragment was eluted with the affinity buffer B (25 mM Tris-HCl at pH 8.0, 0.1 M NaCl, 250 mM imidazole, 0.2 mM $CaCl_2$). Fractions containing the receptor fragment were pooled, concentrated and loaded on a Hiload Superdex 200 gel filtration column using an ÄKTA chromatography system (GE Healthcare). The column was pre-equilibrated and run with 25 mM Tris-HCl at pH 8.0, 0.1 M NaCl, 0.2 mM $CaCl_2$. Dimer and monomer peaks of the *T. ni* TBR3 and CAD fractions were collected separately and concentrated to 1–2 mg ml$^{-1}$. Only TnTBR3 and TnCAD monomers were used for Cry1Ac1 binding studies.

**Fluorescence thermal shift assays.** All assays were performed using a BioRad CFX96 real-time PCR thermal cycler, enabling thermal manipulations and dye fluorescence detection. The fluorescence sensitive dye SYPRO orange (Life Technologies, S6650) was used at a 5× concentration in all assays. The temperature was increased by 0.5 °C each cycle over a temperature range of 25−90 °C. Assay reactions were performed in 96-well white PCR plates (Bio-Rad, number HSP9631), and heat-sealed (Thermo Scientific, number ALPS3000) to reduce volume loss through evaporation. The data were analysed using the CFX manager software.

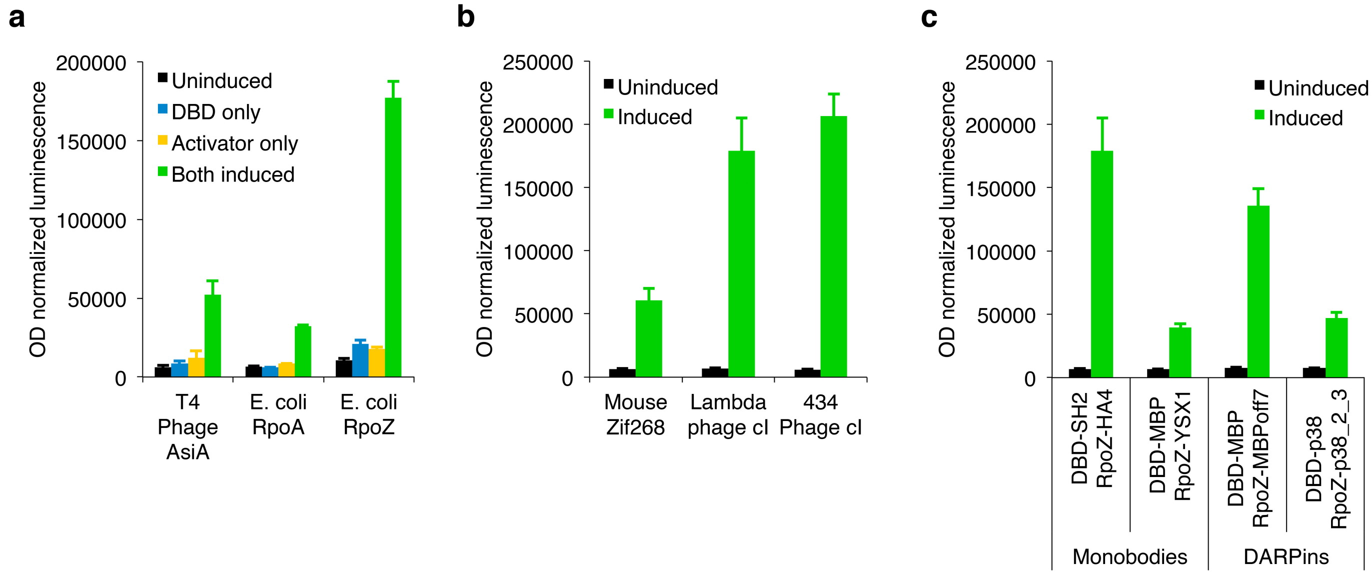
**Protein–protein interaction affinity measurement.** The Octet$^{Qk}$ (ForteBio) and the Dip and Read Ni-NTA (NTA) biosensors were used to measure the affinity of Cry1Ac and its variants to immobilized 6xHis-TnCAD-FL or TnTBR3-FL receptor fragments in 25 mM Tris-HCl at pH 8.5, 0.1 M NaCl, 0.1 mg ml$^{-1}$ BSA, 0.05% Tween 20 according to the manufacturer's instructions. Octet Data Acquisition 7.1.0.100 software was used for data acquisition, and ForteBio Data Analysis 7 software was used for data analysis. At least four readings at different Cry1Ac1 concentrations (2–100 nM) were used for each receptor fragment–Bt toxin interaction and a global fit was used to calculate binding affinities.

**Insect cell assays.** Sf9 cells (Life Technologies) were plated in Sf-900 III SFM (Life Technologies) at a density of 50,000 cells per well in a 96-well optical bottom black plate (Nunc, Thermo Scientific). The cells were incubated at 27 °C overnight to allow for adherence to the plate, and confirmed to be free of mycoplasma contamination using a MycoAlert$^{TM}$ Mycoplasma Detection Kit (Lonza). After overnight incubation, the medium was aspirated from the cells and 100 μl of p3 or p4 generation (third or fourth generation of baculovirus amplification in Sf9 cells after initial transfection with plasmid) recombinant baculovirus encoding each receptor diluted in SFM was added to each well. The plates were kept in a humidified environment to prevent evaporation and incubated at 27 °C for 48 h. Receptor expression was confirmed by western blotting. Toxins were diluted to the same protein concentration in 25 mM sodium carbonate at pH 11, supplemented with 1 mM β-mercaptoethanol, followed by an additional tenfold dilution in unsupplemented Grace's insect media with 2 μM SYTOX green nucleic acid stain (Life Technologies, S7020). The media was removed from the wells without disturbing the attached cells, and the diluted toxins or buffer controls were added to respective wells. The fluorescence was measured on a CLARIOstar microplate reader (BMG Labtech) after incubation for 4 h. The fluorescence intensity of control cells expressing β-glucuronidase (GUS) was subtracted from wells expressing the variable receptor fragments with or without toxins. Replicates were averaged and signal was plotted for each toxin condition.

**Primary insect diet bioassays.** Insect diet bioassays using the evolved consensus Cry1Ac variants were performed as previously described[41]. Briefly, 200 ml of artificial diet in 96-well plates were overlaid with 20 ml aliquots of toxin Bt spore/crystal or Bt crystal suspension, dried, after which wells were infested with neonate insect eggs suspended in 0.2% agar, dried again, sealed with Mylar sheets, and incubated at 20 °C, 60% relative humidity, in complete darkness for 5 days. The plates were scored on day 5 for larval mortality and growth stunting. Each assay was performed in three independent biological replicates with eight insects per replicate.

**Secondary insect diet bioassays.** An inbred Bt-susceptible laboratory strain of *T. ni* (designated the Cornell strain)[42], and a Cry1Ac-resistant strain nearly isogenic to the Cornell strain, GLEN-Cry1Ac-BCS[28], were maintained on a wheat germ-based artificial diet at 27 °C with 50% humidity and a photoperiod of 16 h light and 8 h dark[42]. Diet surface overlay bioassays were conducted to determine the insecticidal activity of the toxins in the susceptible and Cry1Ac-resistant *T. ni*, as previously described[42]. Briefly, 200 μl of each toxin dose solution was spread on the surface of 5 ml of artificial diet in 30-ml plastic rearing cups (diet surface area was ~7 cm$^2$), and ten randomly selected neonatal larvae were placed into each rearing cup after the toxin solution had dried. For each bioassay, seven to eight concentrations of the toxin were used and each treatment included five replicates (50 larvae in total per concentration). Larval growth inhibition (neonates that did not reach second instar after 4 days) and mortality were recorded after 4 days of feeding. The observed larval growth inhibition and mortality were corrected using Abbott's formula[43]. Median inhibitory concentration ($IC_{50}$) and $LC_{50}$ values and their 95% confidence intervals were calculated by probit analysis using the computer program POLO (LeOra Software).


32. Fu, C., Donovan, W. P., Shikapwashya-Hasser, O., Ye, X. & Cole, R. H. Hot Fusion: an efficient method to clone multiple DNA fragments as well as inverted repeats without ligase. *PLoS ONE* **9,** e115318 (2014).
33. Lund, A. M. *et al.* A versatile system for USER cloning-based assembly of expression vectors for mammalian cell engineering. *PLoS ONE* **9,** e96693 (2014).
34. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9,** 357–359 (2012).
35. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
36. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. Preprint at http://adsabs.harvard.edu/abs/2012arXiv1207.3907G%3E (2012).
37. Chaisson, M. J. & Tesler, G. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics* **13,** 238 (2012).
38. Tan, Y. & Donovan, W. P. Deletion of aprA and nprA genes for alkaline protease A and neutral protease A from *Bacillus thuringiensis*: effect on insecticidal crystal proteins. *J. Biotechnol.* **84,** 67–72 (2001).
39. Donovan, W. P. *et al.* Amino acid sequence and entomocidal activity of the P2 crystal protein. An insect toxin from *Bacillus thuringiensis* var. kurstaki. *J. Biol. Chem.* **263,** 561–567 (1988).
40. Nallamsetty, S. *et al.* Efficient site-specific processing of fusion proteins by tobacco vein mottling virus protease in vivo and in vitro. *Protein Expr. Purif.* **38,** 108–115 (2004).
41. Baum, J. A. *et al.* Cotton plants expressing a hemipteran-active *Bacillus thuringiensis* crystal protein impact the development and survival of *Lygus hesperus* (Hemiptera: Miridae) nymphs. *J. Econ. Entomol.* **105,** 616–624 (2012).
42. Kain, W. C. *et al.* Inheritance of resistance to *Bacillus thuringiensis* Cry1Ac toxin in a greenhouse-derived strain of cabbage looper (Lepidoptera: Noctuidae). *J. Econ. Entomol.* **97,** 2073–2078 (2004).
43. Abbott, W. S. A method of computing the effectiveness of an insecticide. 1925. *J. Am. Mosq. Control Assoc.* **3,** 302–303 (1987).

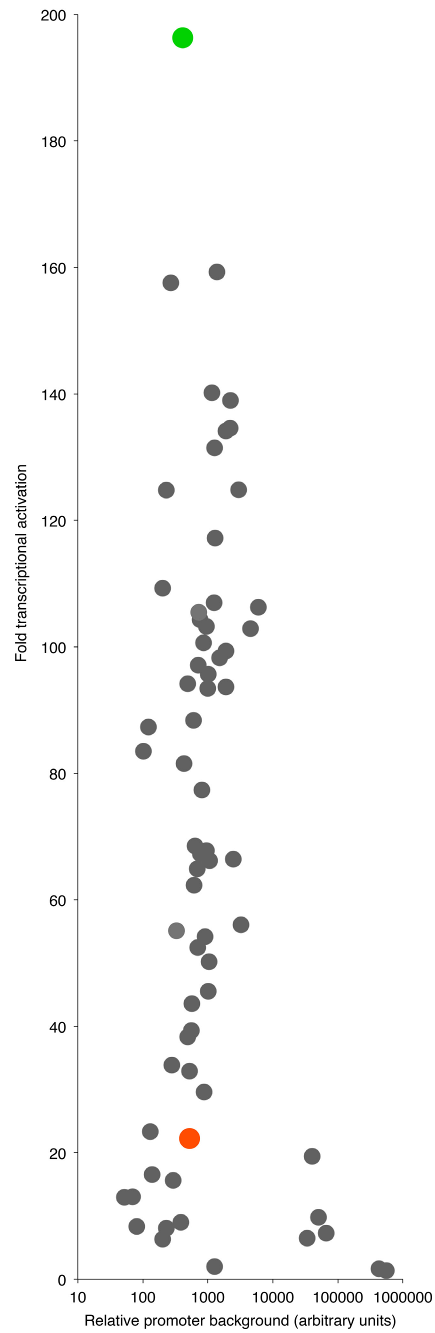



**Extended Data Figure 1 | Bacterial two-hybrid component validation and optimization. a**, Plasmids encoding an IPTG-inducible *λcI–SH2* cassette ('DBD') and an ATc-inducible activator-HA4 cassette ('activator') were co-transformed into the *E. coli* S1030 host strain and induced using either or both small molecules. T4 AsiA-mediated transcriptional activation required low-level expression of the σ70 (R541C/F563Y/L607P) mutant to alleviate AsiA toxicity. Use of RpoZ as the activation domain showed the greatest degree of transcriptional activation (~17-fold). **b**, DNA-binding domain variation shows that multivalent phage repressors yield a greater degree of transcriptional activation than the monomeric zinc finger Zif268. **c**, Transcriptional activation from a combination of the λcI DNA binding domain and RpoZ transcriptional activator was evaluated using several previously evolved protein–protein interactions involving either monobodies or DARPins, showing the generality of binding interaction detection. Error bars, s.d. of at least three independent biological replicates.

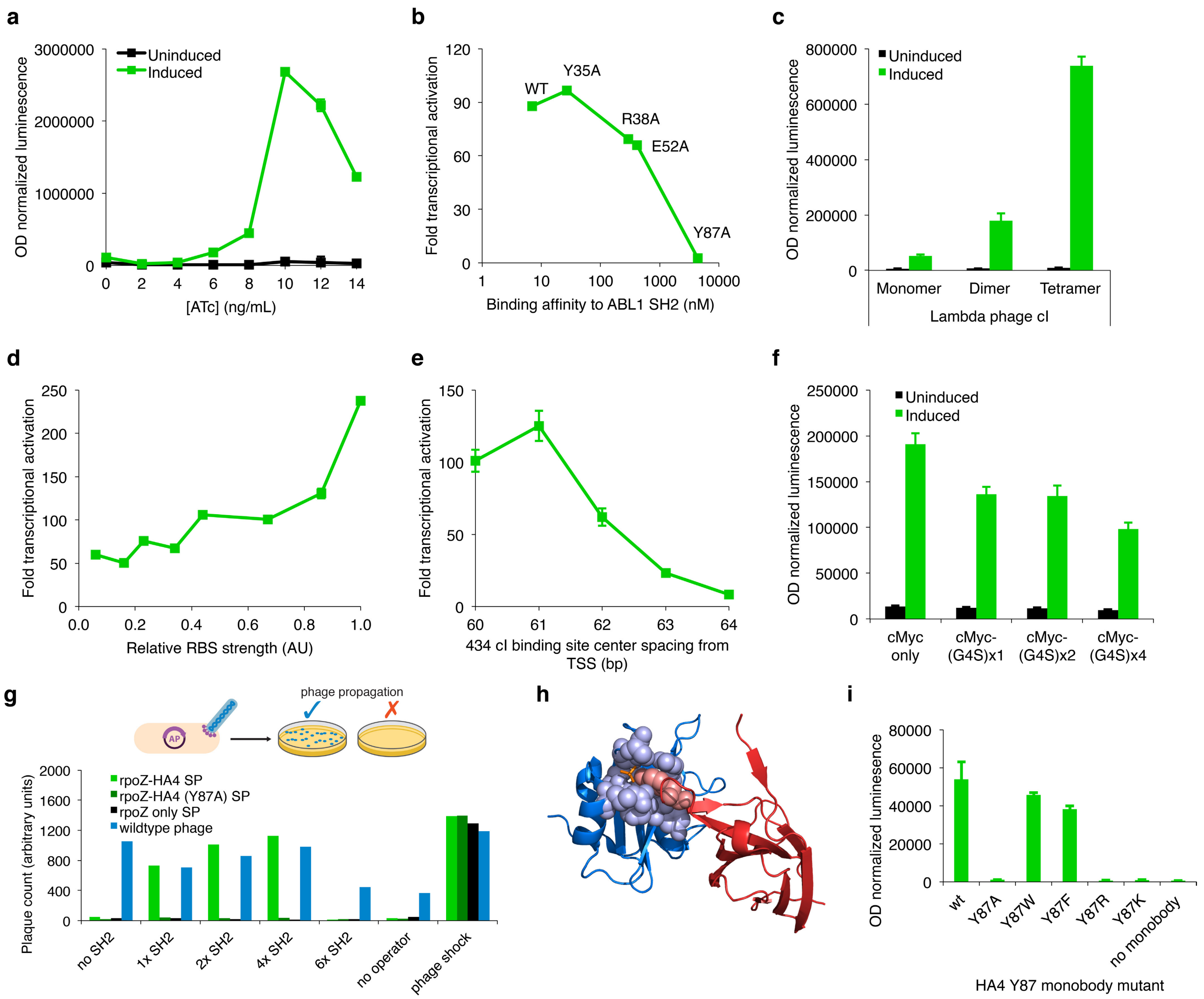
**a**

| DBD | Activator | Promoter | Uninduced | Induced | Fold |
|---|---|---|---|---|---|
| Lambda cI | RpoZ | cattaggcaccccgggctttacactttatgcttccggctcgtatgttgtgtcgaccg | 525 | 11689 | 22 |
| 434 cI | RpoZ | cattaggcaccccgggctttacactttatgcttccggctcgtatgttgtgtcgaccg | 330 | 18167 | 55 |
| 434 cI | RpoZ | cattaggcaccccgggctttacacGtAatgcttccggcGcgtatgttgtgtcgaccg | 3252 | 182235 | 56 |
| 434 cI | RpoZ | cattaggcaccccgggctttacacGtAatgcttccggcGcgtatgGtgtgtcgaccg | 4542 | 467340 | 103 |
| 434 cI | RpoZ | cattaggcaccccgggctttacacGtAatgcttccggcGcgtaCgttgtgtcgaccg | 767 | 51636 | 67 |
| 434 cI | RpoZ | cattaggcaccccgggctttacacGtAatgcttccggcGcgtaCgGtgtgtcgaccg | 945 | 64088 | 68 |
| 434 cI | RpoZ | cattaggcaccccgggcttGacacGtAatgcttccggcGcgtatgttgtgtcgaccg | 432159 | 712599 | 2 |
| 434 cI | RpoZ | cattaggcaccccgggcttGacacGtAatgcttccggcGcgtatgGtgtgtcgaccg | 556230 | 750804 | 1 |
| 434 cI | RpoZ | cattaggcaccccgggctttacGcGtAatgcttccggcGcgtatgttgtgtcgaccg | 1015 | 97100 | 96 |
| 434 cI | RpoZ | cattaggcaccccgggctGtacGcGtAatgcttccggcGcgtatgttgtgtcgaccg | 103 | 8579 | 83 |
| 434 cI | RpoZ | cattaggcaccccgggctGtacGcGtAatgcttccggcGcgtaCgttgtgtcgaccg | 69 | 899 | 13 |
| 434 cI | RpoZ | cattaggcaccccgggctGtacGcGtAatgcttccggcGcgtaCgGtgtgtcgaccg | 52 | 678 | 13 |
| 434 cI | RpoZ | cattaggcaccccgggctGtacGcGtAatgcttccggcGcgtatgttgCgtcgaccg | 202 | 1280 | 6 |
| 434 cI | RpoZ | cattaggcaccccgggctGtacGcGtAatgcttccggcGcgtatgttgGgtcgaccg | 139 | 2307 | 17 |
| 434 cI | RpoZ | cattaggcaccccgggcttGacGcGtAatgcttccggcGcgtatgttgtgtcgaccg | 50526 | 494779 | 10 |
| 434 cI | RpoZ | cattaggcaccccgggcttGacGcGtAatgcttccggcGcgtatgGtgtgtcgaccg | 66275 | 481567 | 7 |
| 434 cI | RpoZ | cattaggcaccccgggcGttacacGtAatgcttccggcGcgtatgGtgtgtcgaccg | 229 | 1851 | 8 |
| 434 cI | RpoZ | cattaggcaccccgggcAttacacGtAatgcttccggcGcgtatgGtgtgtcgaccg | 382 | 3450 | 9 |
| 434 cI | RpoZ | cattaggcaccccgggcCttacacGtAatgcttccggcGcgtatgGtgtgtcgaccg | 131 | 3062 | 23 |
| 434 cI | RpoZ | cattaggcaccccgggctGtacacGtAatgcttccggcGcgtatgGtgtgtcgaccg | 725 | 76430 | 105 |
| 434 cI | RpoZ | cattaggcaccccgggctCtacacGtAatgcttccggcGcgtatgGtgtgtcgaccg | 80 | 668 | 8 |
| 434 cI | RpoZ | cattaggcaccccgggctttacacGtAatgcttccggcGcgtatgttgCgtcgaccg | 859 | 86407 | 101 |
| 434 cI | RpoZ | cattaggcaccccgggctttacacGtAatgcttccggcGcgtatgttgCgtcgaccg | 943 | 97337 | 103 |
| 434 cI | RpoZ | cattaggcaccccgggctttacacGtAatgcttccggcGcgtatgttgCgCcgaccg | 554 | 21807 | 39 |
| 434 cI | RpoZ | cattaggcaccccgggctttacacGtAatgcttccggcGcgtatgttgCgGcgaccg | 428 | 34919 | 82 |
| 434 cI | RpoZ | cattaggcaccccgggctttacacGtAatgcttccggcGcgtatgttgGgtcgaccg | 2229 | 309795 | 139 |
| 434 cI | RpoZ | cattaggcaccccgggctttacacGtAatgcttccggcGcgtatgttgGgCcgaccg | 989 | 92428 | 93 |
| 434 cI | RpoZ | cattaggcaccccgggctttacacGtAatgcttccggcGcgtatgttgGgGcgaccg | 1523 | 149661 | 98 |
| 434 cI | RpoZ | cattaggcaccccgggctttacacGtAatgcttccggcGcgtatgttgAgtcgaccg | 2203 | 296387 | 135 |
| 434 cI | RpoZ | cattaggcaccccgggctttacacGtAatgcttccggcGcgtatgttgAgCcgaccg | 1271 | 167091 | 131 |
| 434 cI | RpoZ | cattaggcaccccgggctttacacGtAatgcttccggcGcgtatgttgAgGcgaccg | 1906 | 189337 | 99 |
| 434 cI | RpoZ | cattaggcaccccgggctttacGcGtAatgcttccggcGcgtatgttgAgCcgaccg | 230 | 28708 | 125 |
| 434 cI | RpoZ | cattaggcaccccgggctttacGcGtAatgcttccggcGcgtatgttgAgGcgaccg | 269 | 42314 | 158 |
| 434 cI | RpoZ | cattaggcaccccgggctGtacacGtAatgcttccggcGcgtatgttgAgCcgaccg | 122 | 10658 | 87 |
| 434 cI | RpoZ | cattaggcaccccgggctGtacacGtAatgcttccggcGcgtatgttgAgGcgaccg | 201 | 22016 | 109 |
| 434 cI | RpoZ | cattaggcaccccgggctttacGcGtAatgcttccggcGcgtatgGtgAgCcgaccg | 566 | 24652 | 44 |
| 434 cI | RpoZ | cattaggcaccccgggctttacGcGtAatgcttccggcGcgtatgGtgAgGcgaccg | 753 | 78485 | 104 |
| 434 cI | RpoZ | cattaggcaccccgggctGtacacGtAatgcttccggcGcgtatgGtgAgCcgaccg | 878 | 26019 | 30 |
| 434 cI | RpoZ | cattaggcaccccgggctGtacacGtAatgcttccggcGcgtatgGtgAgGcgaccg | 1041 | 52307 | 50 |
| 434 cI | RpoZ | cattaggcaccccgggAtttacacGtAatgcttccggcGcgtatgttgtgtcgaccg | 2480 | 164725 | 66 |
| 434 cI | RpoZ | cattaggcaccccgggctttGcacGtAatgcttccggcGcgtatgttgtgtcgaccg | 1015 | 46198 | 46 |
| 434 cI | RpoZ | cattaggcaccccgggctttacTcGtAatgcttccggcGcgtatgttgtgtcgaccg | 1294 | 151683 | 117 |
| 434 cI | RpoZ | cattaggcaccccgggctttacTcGtCatgcttccggcGcgtatgttgtgtcgaccg | 1259 | 134645 | 107 |
| 434 cI | RpoZ | cattaggcaccccgggctttacTcGtAaAgcttccggcGcgtatgttgtgtcgaccg | 412 | 80809 | 196 |
| 434 cI | RpoZ | cattaggcaccccgggctttacTcGtAatgcttccggTGcgtatgttgtgtcgaccg | 613 | 38195 | 62 |
| 434 cI | RpoZ | cattaggcaccccgggctttacCcGtAatgcttccggcGcgtatgttgtgtcgaccg | 601 | 53127 | 88 |
| 434 cI | RpoZ | cattaggcaccccgggctttacacGtCatgcttccggcGcgtatgttgtgtcgaccg | 2975 | 371487 | 125 |
| 434 cI | RpoZ | cattaggcaccccgggctttacacGtCaAgcttccggcGcgtatgttgtgtcgaccg | 1903 | 255198 | 134 |
| 434 cI | RpoZ | cattaggcaccccgggctttacacGtCatgcttccggTGcgtatgttgtgtcgaccg | 1271 | 2597 | 2 |
| 434 cI | RpoZ | cattaggcaccccgggctttacacGtAaAgcttccggcGcgtatgttgtgtcgaccg | 1371 | 218447 | 159 |
| 434 cI | RpoZ | cattaggcaccccgggctttacacGtAaAgcttccggTGcgtatgttgtgtcgaccg | 813 | 62941 | 77 |
| 434 cI | RpoZ | cattaggcaccccgggctttacacGtAatgcCtccggcGcgtatgttgtgtcgaccg | 33730 | 218546 | 6 |
| 434 cI | RpoZ | cattaggcaccccgggctttacacGtAatgcttccggTGcgtatgttgtgtcgaccg | 1146 | 160609 | 140 |
| 434 cI | RpoZ | cattaggcaccccgggctttacacGtAatgcttccggcGTgtatgttgtgtcgaccg | 39981 | 777784 | 19 |
| 434 cI | RpoZ | cattaggcaccccgggctttacacGtAatgcttccggcGAgtatgttgtgtcgaccg | 6010 | 638749 | 106 |
| 434 cI | RpoZ | cattaggcaccccgggctttacacGtAatgcttccggcGcgtaGgttgtgtcgaccg | 489 | 46068 | 94 |
| 434 cI | RpoZ | cattaggcaccccgggctttacacGtAatgcttccggcGcgtatgtAgtgtcgaccg | 291 | 4553 | 16 |
| 434 cI | RpoZ | cattaggcaccccgggctttacTcGtAatgcttccggcGcgtatgttgAgCcgaccg | 685 | 44500 | 65 |
| 434 cI | RpoZ | cattaggcaccccgggctttacTcGtCatgcttccggcGcgtatgttgAgCcgaccg | 638 | 43688 | 68 |
| 434 cI | RpoZ | cattaggcaccccgggctttacTcGtAaAgcttccggcGcgtatgttgAgCcgaccg | 488 | 18713 | 38 |
| 434 cI | RpoZ | cattaggcaccccgggctttacTcGtAatgcttccggTGcgtatgttgAgCcgaccg | 278 | 9431 | 34 |
| 434 cI | RpoZ | cattaggcaccccgggctttacacGtCatgcttccggcGcgtatgttgAgCcgaccg | 1890 | 176933 | 94 |
| 434 cI | RpoZ | cattaggcaccccgggctttacacGtCaAgcttccggcGcgtatgttgAgCcgaccg | 709 | 68843 | 97 |
| 434 cI | RpoZ | cattaggcaccccgggctttacacGtCatgcttccggTGcgtatgttgAgCcgaccg | 908 | 49205 | 54 |
| 434 cI | RpoZ | cattaggcaccccgggctttacacGtAaAgcttccggcGcgtatgttgAgCcgaccg | 1060 | 70184 | 66 |
| 434 cI | RpoZ | cattaggcaccccgggctttacacGtAaAgcttccggTGcgtatgttgAgCcgaccg | 524 | 17228 | 33 |
| 434 cI | RpoZ | cattaggcaccccgggctttacacGtAatgcttccggTGcgtatgttgAgCcgaccg | 696 | 36530 | 52 |

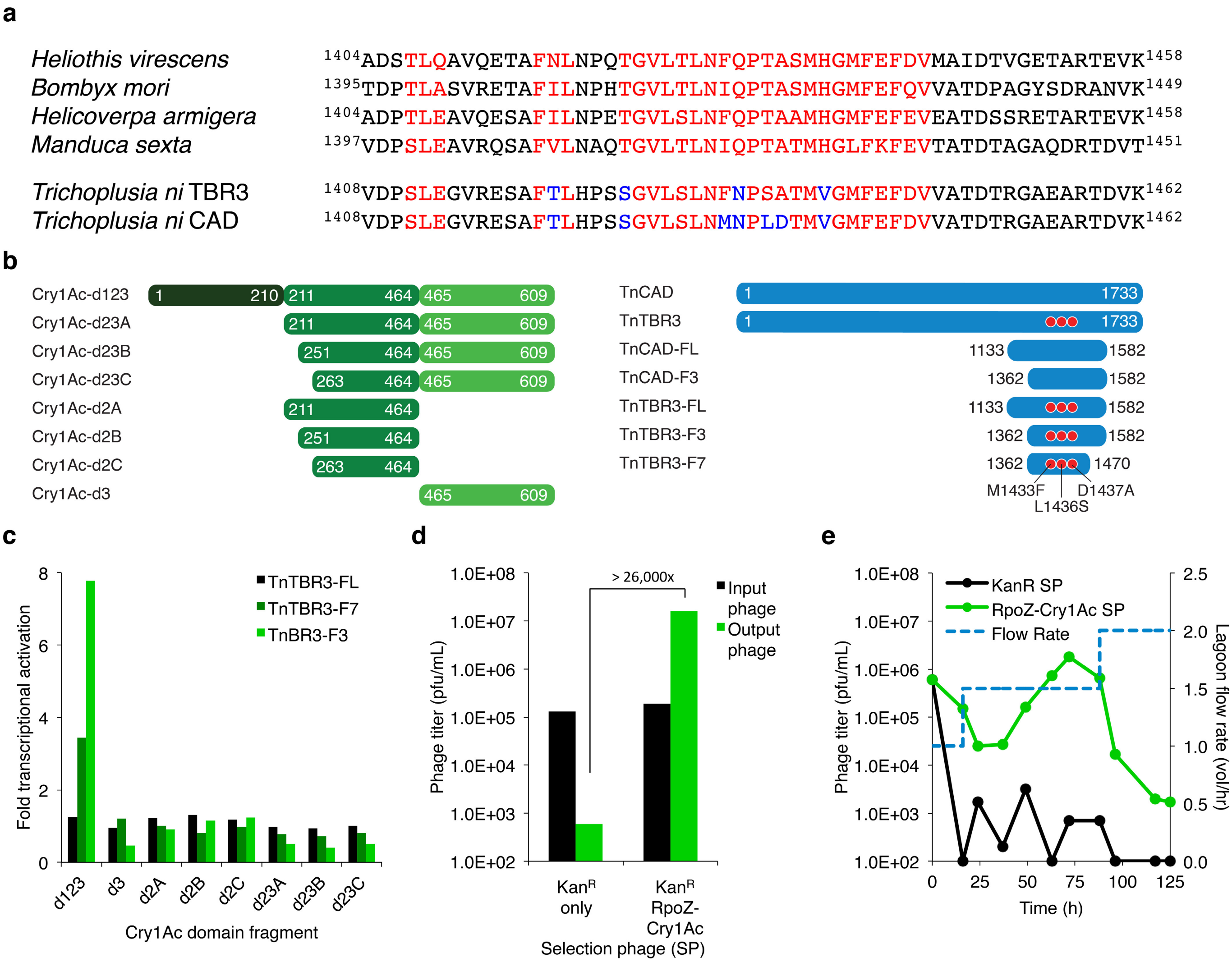
**b**



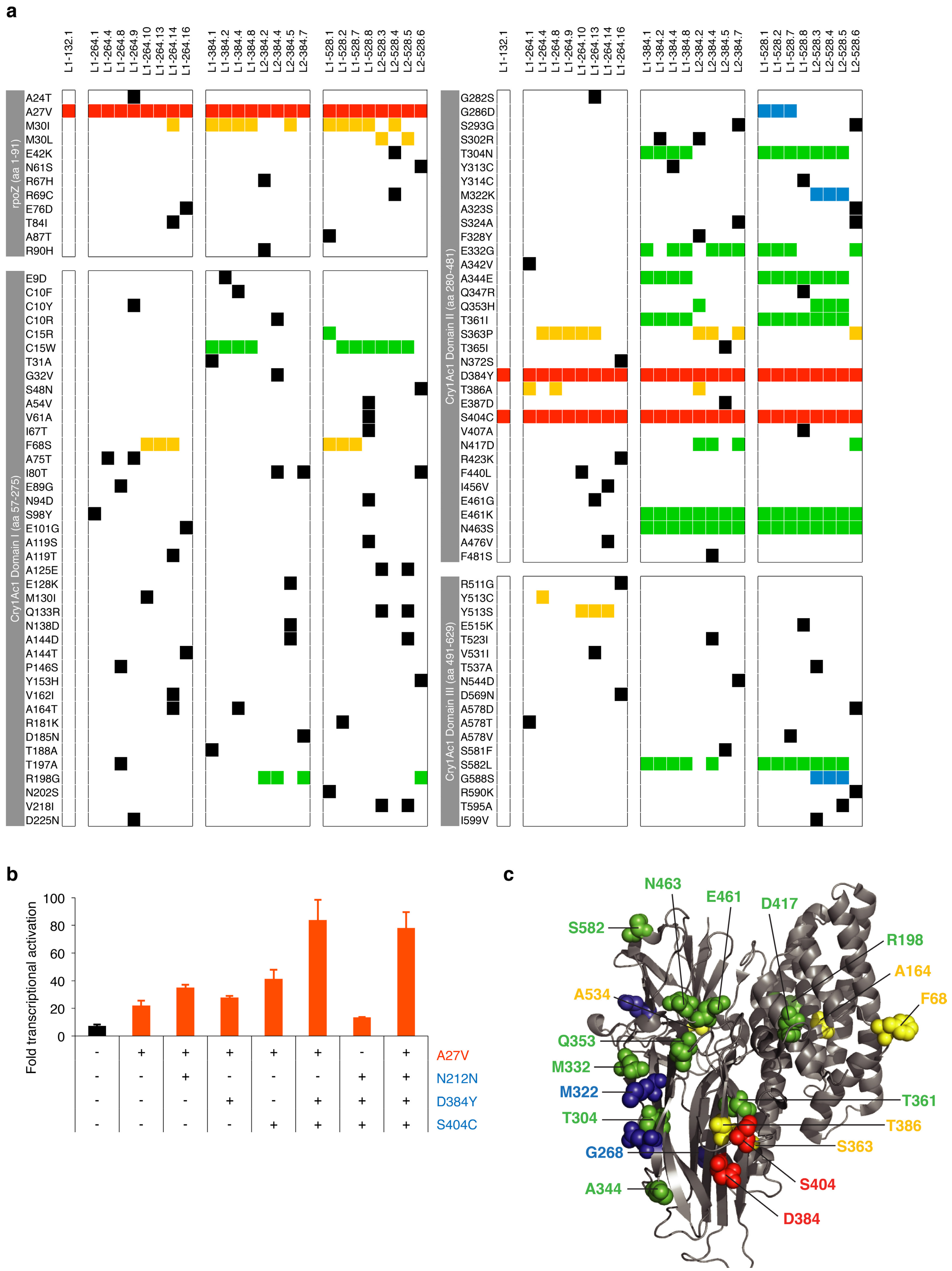

**Extended Data Figure 2 | Optimization of the $P_{lacZ}$ promoter for improved sensitivity and dynamic range. a**, Promoter and DNA-binding domain combinations tested during $P_{lacZ}$ optimization, showing uninduced and induced levels of absorbance-normalized luminescence. The SH2/HA4 interaction pair was used in all cases. The fold activation in each case was calculated as the ratio of the induced and uninduced luciferase expression signals. **b**, Graphical representation of the data in **a**, showing the wide distribution of promoter background levels and degrees of transcriptional activation. In **a** and **b**, the red and green dots indicate the starting ($P_{lac62}$) and final ($P_{lacZ\text{-}opt}$) promoter/DNA-binding domain combinations, respectively. Each data point in **b** reflects the average of at least three independent biological replicates.



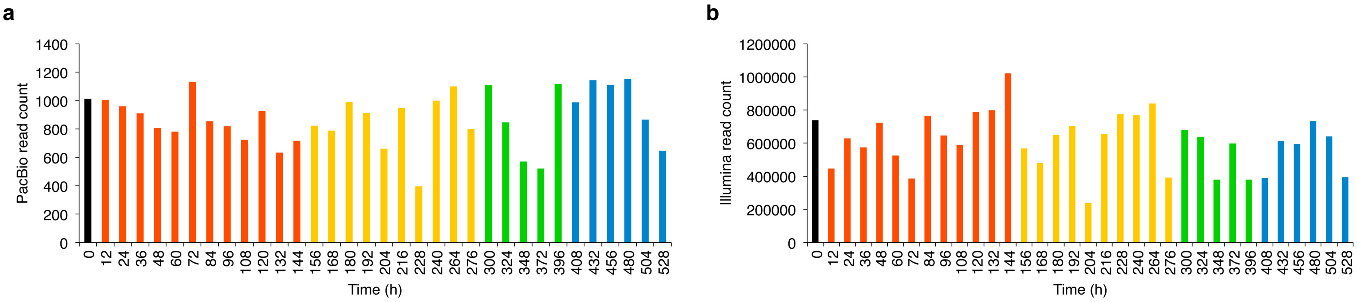

**Extended Data Figure 3 | Bacterial two-hybrid optimization. a**, Inducer titration of the interacting fusion proteins driving the two-hyrbid system. The black and green lines represent the uninduced (0 μM IPTG) and induced (1 μM IPTG) levels of IPTG-inducible *434cI–SH2* expression, while ATc induces expression of the *rpoZ–HA4* cassette. In subsequent graphs and assays, the expression level resulting from the IPTG-inducible $P_{lac}$ promoter was measured by western blot and approximated using a constitutive promoter to reduce experimental variability. **b**, Degree of transcriptional activation using HA4 monobody mutants correlated with known binding affinities. The highest levels of activation resulted from $K_d$ = low nanomolar affinities, while weak affinities in the $K_d$ = low micromolar range could still be detected. **c**, Relationship between DNA-binding domain multivalency state (monomeric, dimeric, or tetrameric DNA-binding domain fused to the SH2 domain) and transcriptional activation resulting from the SH2/HA4 interaction, with higher multivalency states yielding greater activation levels. **d**, RBS modification enables robust modulation of the relative activation levels from the $P_{lacZ\text{-}opt}$ promoter using the SH2/HA4 interaction. **e**, Operator–promoter binding site spacing strongly affects transcriptional activation levels; 434cI binding at 61 base pairs upstream of the $P_{lacZ\text{-}opt}$ promoter resulted in the most robust activation. **f**, Linker extension to include one, two, or three $G_4S$ motifs result in reduced activation levels using the SH2/HA4 interacting pair. **g**, Phage plaque formation as a function of target protein multivalency. 'No operator' indicates a scrambled 434cI operator control accessory plasmid; 'phage control' indicates an accessory plasmid in which the phage shock promoter (activated by phage infection) drives gene III expression. **h**, Co-crystal structure of the ABL1 SH2 (blue) bound to the HA4 monobody (red), highlighting the interaction of HA4 Y87 (red spheres) with key residues of the phosphotyrosine-binding pocket (blue spheres) of the SH2 domain (Protein Data Bank accession number 3K2M). The phosphate ion is shown in orange at the interaction interface. **i**, Apparent binding activity of mutants of the HA4 monobody at position 87. Tyrosine, tryptophan, and phenylalanine are tolerated at position 87 and enable protein–protein interaction by bacterial two-hybrid assay. Error bars, s.d. of at least three independent biological replicates.

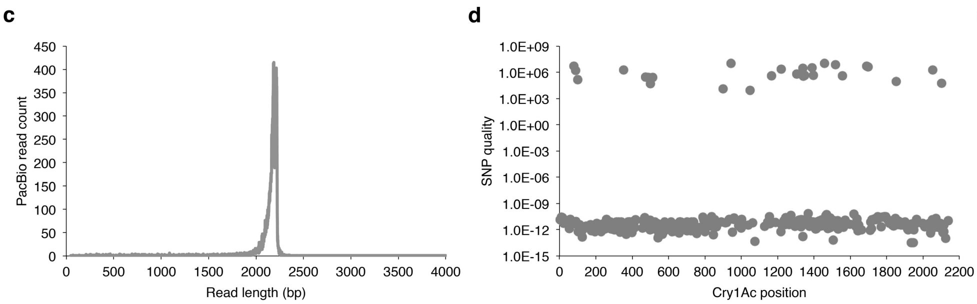



**Extended Data Figure 4 | Choice of Cry1Ac and TnTBR3 fragments used in PACE. a,** Protein sequence alignment of known Cry1Ac-binding motifs from cadherin receptors in several lepidopteran species, as well as the cadherin receptor from *T. ni* (TnCAD). The toxin-binding region (TBR; shown in red) of the known Cry1Ac-binding motifs differs from TnCAD at seven positions (shown in blue). Mutation of three residues in the TnCAD TBR (M1433F, L1436S, and D1437A) to resemble the corresponding positions of the cadherin-receptor TBRs yielded the evolutionary stepping-stone target TnTBR3. **b,** Schematic representations of the Cry1Ac and *T. ni* TBR3/CAD full-length receptors and fragments tested in this study. The red stars in the TnTBR3 variants represent the three mutations introduced into TnCAD to generate TnTBR3. **c,** Transcriptional activation assay using Cry1Ac and TnTBR3 fragments shows that the greatest degree of transcriptional activation resulted from full-length Cry1Ac together with TBR3 fragment 3 (TnTBR3-F3). RpoZ–Cry1Ac and 434cI–TnTBR3 fusions were used in all cases. **d,** Overnight phage enrichment assays using selection phages that encode either kanamycin resistance (Kan$^R$) only or Kan$^R$ together with RpoZ–Cry1Ac. Compared with the Kan$^R$-only selection phage, the RpoZ–Cry1Ac selection phage enriches >26,000-fold overnight. **e,** Continuous propagation assays in the PACE format using either the Kan$^R$-only selection phage or the RpoZ–Cry1Ac selection phage show that the moderate affinity of Cry1Ac for TnTBR3 allows phage propagation at low flow rates (≤1.5 lagoon volumes per hour).



**Extended Data Figure 5 | Single-clone sequencing and evolved Cry1Ac characterization after PACE using the bacterial two-hybrid luminescence reporter. a**, Coding mutations of the tested RpoZ–Cry1Ac clones at the end of each of the four segments of PACE. Consensus mutations are coloured according to the segment in which they became highly enriched in the population (Fig. 3a). Mutations coloured in black were observed at low abundance (≤5% of sequenced clones). **b**, Mutational dissection of the consensus mutations from the first segment of PACE reveals the requirement for both D384Y and S404C to achieve high-level transcriptional activation using the TnTBR3-F3 target. Mutations listed in red occurred in the RpoZ activation domain, whereas mutations listed in blue occurred in the Cry1Ac domain. Error bars, s.d. of at least three independent biological replicates. **c**, Structure of wild-type Cry1Ac (Protein Data Bank accession number 4ARX) showing the positions of the evolved consensus mutations. The colours correspond to the PACE segments shown in Fig. 3 during which the mutations became highly abundant.

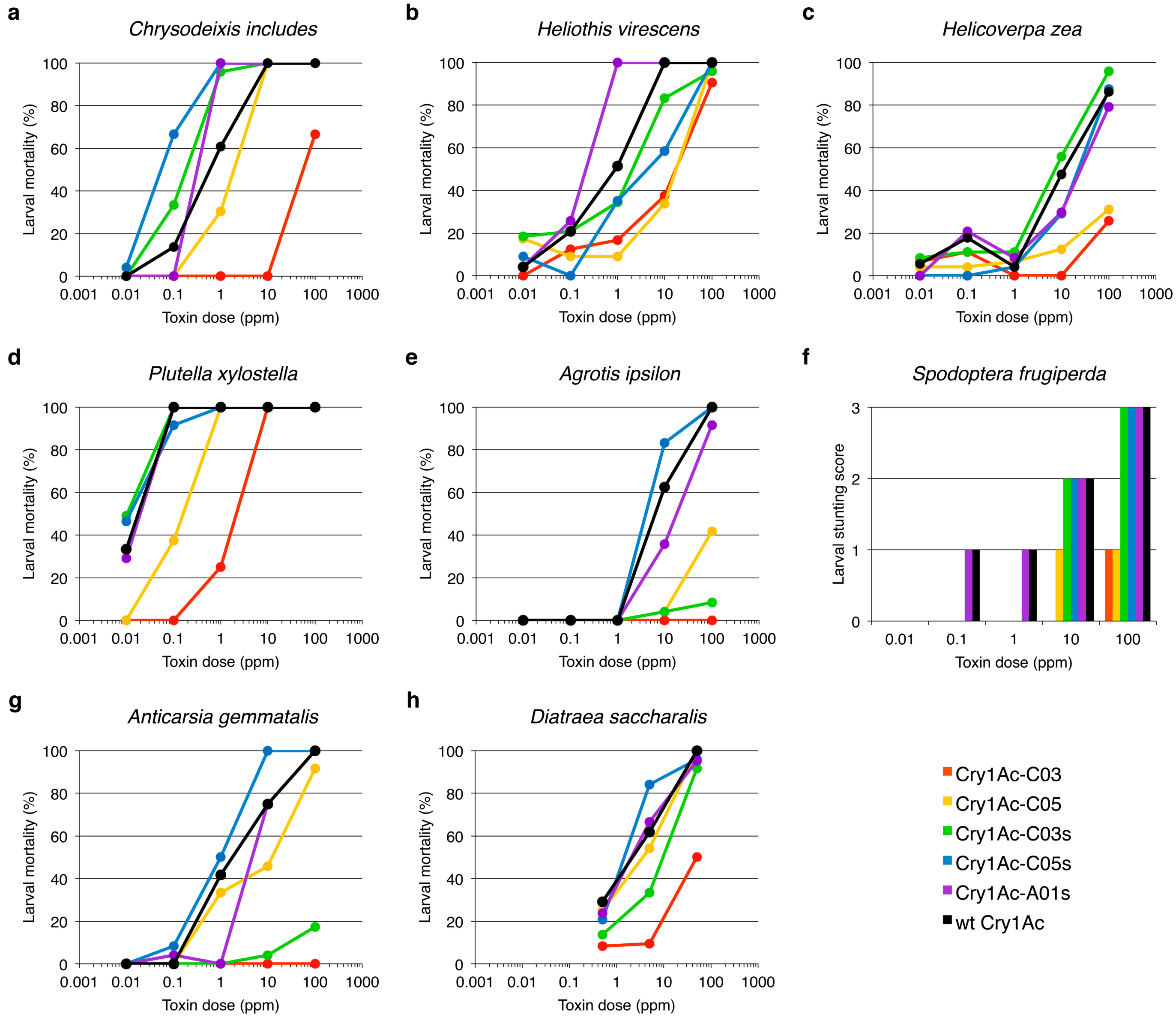



**e**

| Gene | Position | Ref | Allele | Mutation | Synonymous |
|---|---|---|---|---|---|
| *rpoZ* | 80 | C | T | A-76V | N |
| | 90 | G | A | M-73I | N |
| | 102 | A | G | G-69G | S |
| *cry1Ac* | 354 | T | G | C15W | N |
| | 474 | A | G | G55G | S |
| | 501 | A | C | I64I | S |
| | 512 | T | C | F68S | N |
| | 901 | A | G | R198G | N |
| | 945 | T | C | N212N | S |
| | 1047 | G | A | L246L | S |
| | 1166 | G | A | G286D | N |
| | 1220 | C | A | T304N | N |
| | 1304 | A | G | E332G | N |
| | 1340 | C | A | A344E | N |
| | 1344 | T | C | A345A | S |
| | 1349 | A | G | Q347R | N |
| | 1391 | C | T | T361I | N |
| | 1396 | T | C | S363P | N |
| | 1459 | G | T | D384Y | N |
| | 1519 | A | T | S404C | N |
| | 1558 | A | G | N417D | N |
| | 1690 | G | A | E461K | N |
| | 1697 | A | G | N463S | N |
| | 1852 | G | A | E515K | N |
| | 2054 | C | T | S582L | N |

**Extended Data Figure 6 | High-throughput DNA sequencing of PACE Cry1Ac selection phage libraries.** The number of reads mapped to the wild-type *rpoZ–Cry1Ac* reference sequence using (**a**) Pacific Biosciences (PacBio) or (**b**) Illumina sequencing. Time points are coloured according to the corresponding segment of the PACE experiment (Fig. 3a). **c**, In general, most PacBio reads aligned to the wt *rpoZ–Cry1Ac* reference sequence were found to cluster around ~2,200 base pairs, corresponding to the size of the full-length fusion gene and indicating high-quality sequencing reads. **d**, Illumina high-throughput sequencing yielded several high-quality single nucleotide polymorphisms across all time points. The corresponding mutations are shown in **e**.




**Extended Data Figure 7 | Insect diet bioassay activity of PACE-evolved Cry1Ac variants against various agricultural pests.** Two consensus and three stabilized PACE-evolved Cry1Ac variants were tested for activity in eleven pests: **a**, *C. includes* (soybean looper); **b**, *Heliothis virescens* (tobacco budworm); **c**, *Helicoverpa zea* (corn earworm); **d**, *Plutella xylostella* (diamondback moth); **e**, *Agrotis ipsilon* (black cutworm); **f**, *Spodoptera frugiperda* (fall armyworm); **g**, *Anticarsia gemmatalis* (velvetbean caterpillar); **h**, *Diatraea saccharalis* (sugarcane borer); *Spodoptera eridania* (southern armyworm); *Leptinotarsa decemlineata* (Colorado potato beetle); and *Lygus lineolaris* (tarnished plant bug). Stabilized variants showed enhanced activity in *C. includens* and *H. virescens* compared with wild-type Cry1Ac, and comparable activity to wild-type Cry1Ac in *H. zea*, *P. xylostella*, *A. ipsilon*, *S. frugiperda*, *A. gemmatalis*, and *D. saccharalis*. No activity was observed for any of the Cry1Ac variants at any tested dose for *S. eridania*, *L. decemlineata*, or *L. lineolaris*. No insect larvae mortality was observed for *S. frugiperda*, although high toxin doses greatly stunted growth.

**a**

| | *T. ni* | *C. includens* | *H. virescens* | *H. zea* | *P. xylostella* | *A. ipsilon* | *S. frugiperda* | *L. decemlineata* | *D. virgifera* |
|---|---|---|---|---|---|---|---|---|---|
| *T. ni* | - | 78.6 (1362) | 57.9 (1004) | 57.3 (993) | 53.1 (920) | 51.8 (897) | 51.2 (888) | 25.6 (443) | 26.1 (453) |
| *C. includens* | 76.8 (1362) | - | 55.7 (988) | 54 (958) | 51.8 (918) | 49.9 (884) | 49.8 (883) | 24.6 (437) | 25.5 (452) |
| *H. virescens* | 58 (1004) | 57 (988) | - | 84.5 (1464) | 55.4 (959) | 58.9 (1021) | 57.3 (992) | 26 (451) | 25.7 (445) |
| *H. zea* | 57.4 (993) | 55.4 (958) | 84.6 (1464) | - | 54.6 (945) | 58.7 (1016) | 57.3 (992) | 26.3 (455) | 25.4 (439) |
| *P. xylostella* | 53.6 (920) | 53.5 (918) | 55.9 (959) | 55 (945) | - | 49.4 (848) | 50.5 (867) | 24.6 (422) | 25.3 (435) |
| *A. ipsilon* | 51 (897) | 50.2 (884) | 58 (1021) | 57.7 (1016) | 48.2 (848) | - | 53.1 (935) | 24 (422) | 25 (440) |
| *S. frugiperda* | 51.2 (888) | 50.9 (883) | 57.2 (992) | 57.2 (992) | 50 (867) | 53.9 (935) | - | 25.7 (445) | 25.1 (436) |
| *L. decemlineata* | 26.3 (443) | 25.9 (437) | 26.7 (451) | 27 (455) | 25 (422) | 25 (422) | 26.4 (445) | - | 39.6 (668) |
| *D. virgifera* | 26.8 (453) | 26.8 (452) | 26.4 (445) | 26 (439) | 25.8 (435) | 26.1 (440) | 25.8 (436) | 39.6 (668) | - |

< 40%

< 70%

> 70%

**b**

| | *T. ni* | *C. includens* | *H. virescens* | *H. zea* | *P. xylostella* | *A. ipsilon* | *S. frugiperda* | *L. decemlineata* | *D. virgifera* |
|---|---|---|---|---|---|---|---|---|---|
| *T. ni* | - | 76.9 (170) | 55.2 (122) | 54.3 (120) | 52 (115) | 47.5 (105) | 47.1 (104) | 24 (53) | 22.2 (49) |
| *C. includens* | 75.2 (170) | - | 51.3 (116) | 51.3 (116) | 47.8 (108) | 42.5 (96) | 42.5 (96) | 22.6 (51) | 22.6 (51) |
| *H. virescens* | 55 (122) | 52.3 (116) | - | 81.5 (181) | 50 (111) | 49.5 (110) | 47.7 (106) | 23.4 (52) | 19.8 (44) |
| *H. zea* | 53.8 (120) | 52 (116) | 81.2 (181) | - | 49.3 (110) | 50.2 (112) | 47.5 (106) | 24.2 (54) | 19.3 (43) |
| *P. xylostella* | 52 (115) | 48.9 (108) | 50.2 (111) | 49.8 (110) | - | 43.9 (97) | 48 (106) | 22.2 (49) | 19 (42) |
| *A. ipsilon* | 47.3 (105) | 43.2 (96) | 49.5 (110) | 50.5 (112) | 43.7 (97) | - | 42.8 (95) | 18.5 (41) | 18.9 (42) |
| *S. frugiperda* | 46.8 (104) | 43.2 (96) | 47.7 (106) | 47.7 (106) | 47.7 (106) | 42.8 (95) | - | 24.8 (55) | 23.4 (52) |
| *L. decemlineata* | 22.3 (53) | 21.4 (51) | 21.8 (52) | 22.7 (54) | 20.6 (49) | 17.2 (41) | 23.1 (55) | - | 38.2 (91) |
| *D. virgifera* | 20.2 (49) | 21 (51) | 18.1 (44) | 17.7 (43) | 17.3 (42) | 17.3 (42) | 21.4 (52) | 37.4 (91) | - |

**Extended Data Figure 8 | Comparison of cadherin receptor sequence identity.** The percentage sequence identity using the full-length cadherin receptor (**a**) or fragment used for directed evolution experiments (**b**) for insects tested in Extended Data Fig. 7. Numbers in parentheses denote the number of identical amino acids between the two receptors. In general, mortality and stunting data from diet bioassays correlate with cadherin receptor sequence identity.

**Extended Data Table 1 | Insect bioassays against susceptible and resistant *T. ni***

| | | Toxin | $LC_{50}$ (ppm) | 95% CL | Slope | SE | Relative potency (%) |
|---|---|---|---|---|---|---|---|
| Susceptible *T. ni* | Mortality | Cry1Ac | 0.039 | 0.019 - 0.069 | 2.54 | 0.26 | 100 |
| | | Cry1Ac-C03 | 0.793 | 0.505 - 1.082 | 2.84 | 0.41 | 5 |
| | | Cry1Ac-C05 | 0.715 | 0.407 - 1.176 | 1.78 | 0.22 | 5 |
| | | Cry1Ac-C03s | 0.018 | 0.014 - 0.020 | 4.68 | 0.75 | 217 |
| | | Cry1Ac-C05s | 0.035 | 0.026 - 0.045 | 3.59 | 0.41 | 111 |
| | | Cry1Ac-A01s | 0.021 | 0.015 - 0.024 | 4.82 | 1.09 | 186 |
| | | Toxin | $IC_{50}$ (ppm) | 95% CL | Slope | SE | Relative potency (%) |
| | Growth inhibition | Cry1Ac | 0.019 | 0.011 - 0.027 | 3.09 | 0.39 | 100 |
| | | Cry1Ac-C03 | 0.136 | 0.110 - 0.160 | 4.00 | 0.62 | 14 |
| | | Cry1Ac-C05 | 0.217 | 0.167 - 0.268 | 2.59 | 0.32 | 9 |
| | | Cry1Ac-C03s | 0.007 | 0.003 - 0.010 | 3.65 | 0.61 | 271 |
| | | Cry1Ac-C05s | 0.016 | 0.014 - 0.018 | 5.53 | 0.82 | 119 |
| | | Cry1Ac-A01s | 0.005 | 0.004 - 0.006 | 4.92 | 0.9 | 380 |
| | | Toxin | $LC_{50}$ (ppm) | 95% CL | Slope | SE | Relative potency (%) |
| Resistant *T. ni* | Mortality | Cry1Ac | 51.229 | 9.929 - 90.241 | 1.89 | 0.36 | 100 |
| | | Cry1Ac-C03 | 408.713 | 263.629 - 680.973 | 0.81 | 0.1 | 13 |
| | | Cry1Ac-C05 | 235.698 | 79.467 - 510.323 | 1.12 | 0.15 | 22 |
| | | Cry1Ac-C03s | 1.841 | 1.390 - 2.312 | 2.25 | 0.28 | 2783 |
| | | Cry1Ac-C05s | 1.938 | 1.550 - 2.352 | 2.55 | 0.29 | 2643 |
| | | Cry1Ac-A01s | 0.153 | 0.046 - 0.289 | 2.01 | 0.22 | 33483 |
| | | Toxin | $IC_{50}$ (ppm) | 95% CL | Slope | SE | Relative potency (%) |
| | Growth inhibition | Cry1Ac | 23.402 | 4.587 - 46.512 | 1.49 | 0.25 | 100 |
| | | Cry1Ac-C03 | 56.626 | 40.600 - 75.685 | 1.84 | 0.21 | 41 |
| | | Cry1Ac-C05 | 47.232 | 20.236 - 90.729 | 1.16 | 0.12 | 50 |
| | | Cry1Ac-C03s | 0.733 | 0.515 - 0.949 | 2.06 | 0.28 | 3193 |
| | | Cry1Ac-C05s | 1.116 | 0.797 - 1.484 | 2.19 | 0.23 | 2097 |
| | | Cry1Ac-A01s | 0.083 | 0.061 - 0.104 | 2.57 | 0.38 | 28195 |

The $LC_{50}$ and $IC_{50}$ values were determined using seven to eight concentrations of the indicated toxins in an insect diet surface overlay bioassay using either Cry1Ac-susceptible or Cry1Ac-resistant *T. ni* neonatal larvae. Each toxin concentration was tested in five replicates, each of which contained ten randomly selected neonatal larvae. In each case, the 95% confidence interval (95% CI), the slope of the best fit (Slope) and the standard error (s.e.) is given. The relative potency (%) has been normalized to the activity of wild-type Cry1Ac for each case.

**Extended Data Table 2 | Plasmids used in this work**

| Plasmid name | Class (resistance) | Origin of replication | ORF 1 | | ORF 2 | | ORF 3 | | Figure(s) |
|---|---|---|---|---|---|---|---|---|---|
| | | | Promoter | Gene | Promoter | Gene | Promoter | Gene | |
| pAB029f | CP (spec$^{R}$) | ColEI | $P_{lacZ}$ | λcI-SH2$_{ABL1}$ | $P_{lacIq}$ | lacI | - | - | S1A, S1C |
| pAB030g | AP (carb$^{R}$) | SC101 | $P_{lac62}$ | gIII, luxAB | $P_{tet}$ | rpoA-HA4 | - | - | S1A |
| pAB031a | CP (spec$^{R}$) | ColEI | $P_{lacZ}$ | HA4-Zif268 | $P_{lacIq}$ | lacI | - | - | S1B |
| pAB035a | AP (carb$^{R}$) | SC101 | $P_{lac62}$ | gIII, luxAB | $P_{tet}$ | rpoZ-SH2$_{ABL1}$ | - | - | S1B |
| pAB035h | AP (carb$^{R}$) | SC101 | $P_{lac62}$ | gIII, luxAB | $P_{tet}$ | rpoZ-HA4 | - | - | S1A, S1C |
| pAB042a | AP (carb$^{R}$) | SC101 | $P_{lac62}$ | gIII, luxAB | $P_{tet}$ | rpoZ-YSX1 | - | - | S1C |
| pAB042b | AP (carb$^{R}$) | SC101 | $P_{lac62}$ | gIII, luxAB | $P_{tet}$ | rpoZ-MBPoff7 | - | - | S1C |
| pAB042d | AP (carb$^{R}$) | SC101 | $P_{lac62}$ | gIII, luxAB | $P_{tet}$ | rpoZ-p38_2_3 | - | - | S1C |
| pAB043a | CP (spec$^{R}$) | ColEI | $P_{lacZ}$ | λcI-MBP | $P_{lacIq}$ | lacI | - | - | S1C |
| pAB043c | CP (spec$^{R}$) | ColEI | $P_{lacZ}$ | λcI-p38α | $P_{lacIq}$ | lacI | - | - | S1C |
| pAB045a | CP (spec$^{R}$) | ColEI | $P_{lacZ}$ | rpoZ-HA4 | $P_{lacIq}$ | lacI | - | - | S1B, S3A, S3C |
| pAB049b | CP (spec$^{R}$) | ColEI | $P_{lacZ}$ | HA4-AsiA | $P_{lacIq}$ | lacI | - | - | S1A |
| pAB051e | AP (carb$^{R}$) | SC101 | $P_{Ext\text{-}10cons\lambda op\text{-}62}$ | gIII, luxAB | $P_{tet}$ | λcI-SH2$_{ABL1}$ | $P_{lpp}$ | rpoD (R541C/F563Y/L607P) | S1A |
| pAB060c | AP (carb$^{R}$) | SC101 | $P_{lac62}$ | luxAB | $P_{pro1}$ | λcI-SH2$_{ABL1}$ | - | - | S1B |
| pAB061i | CP (spec$^{R}$) | ColEI | $P_{pro1}$ | rpoZ-HA4 | $P_{lacIq}$ | lacI | - | - | 2A, S4B |
| pAB061i10 | CP (spec$^{R}$) | ColEI | $P_{pro1}$ | rpoZ-HA4 (Y87R) | $P_{lacIq}$ | lacI | - | - | S4B |
| pAB061i11 | CP (spec$^{R}$) | ColEI | $P_{pro1}$ | rpoZ-HA4 (Y87K) | $P_{lacIq}$ | lacI | - | - | S4B |
| pAB061i2 | CP (spec$^{R}$) | ColEI | $P_{pro1}$ | rpoZ-HA4 (Y87A) | $P_{lacIq}$ | lacI | - | - | 2A, S4B |
| pAB061i3 | CP (spec$^{R}$) | ColEI | $P_{pro1}$ | rpoZ | $P_{lacIq}$ | lacI | - | - | 2A, S4B |
| pAB061i8 | CP (spec$^{R}$) | ColEI | $P_{pro1}$ | rpoZ-HA4 (Y87W) | $P_{lacIq}$ | lacI | - | - | S4B |
| pAB061i9 | CP (spec$^{R}$) | ColEI | $P_{pro1}$ | rpoZ-HA4 (Y87F) | $P_{lacIq}$ | lacI | - | - | S4B |
| pAB064d | AP (carb$^{R}$) | SC101 | $P_{lac62}$ | luxAB | $P_{pro1}$ | 434cI-SH2$_{ABL1}$ | - | - | S1B |
| pAB076i3 | AP (carb$^{R}$) | SC101 | $P_{lacZ\text{-}opt}$ (OR1) | gIII, luxAB | $P_{pro1}$ | 434cI-SH2$_{ABL1}$ | - | - | 2B |
| pAB076i5 | AP (carb$^{R}$) | SC101 | $P_{lacZ\text{-}opt}$ (OR1+2+3) | gIII, luxAB | $P_{pro1}$ | 434cI-SH2$_{ABL1}$ | - | - | 2B |
| pAB076i6 | AP (carb$^{R}$) | SC101 | $P_{lacZ\text{-}opt}$ (no operator) | gIII, luxAB | $P_{pro1}$ | 434cI-SH2$_{ABL1}$ | - | - | 2B |
| pAB076i7 | AP (carb$^{R}$) | SC101 | $P_{lacZ\text{-}opt}$ (OR1) | gIII, luxAB | $P_{pro1}$ | 434cI(RR69)-SH2$_{ABL1}$ | - | - | 2B |
| pAB078d10 | AP (carb$^{R}$) | SC101 | $P_{lacZ\text{-}opt}$ (OR1) | luxAB | $P_{pro1}$ | 434cI(RR69)-SH2$_{ABL1}$ | - | - | 2A, S3C |
| pAB078d3 | AP (carb$^{R}$) | SC101 | $P_{lacZ\text{-}opt}$ (OR1) | luxAB | $P_{pro1}$ | 434cI-SH2$_{ABL1}$ | - | - | 2A, S3B, S3C, S4B |
| pAB078d5 | AP (carb$^{R}$) | SC101 | $P_{lacZ\text{-}opt}$ (OR1) | luxAB | $P_{pro1}$ | 434cI | - | - | 2A |
| pAB078d7 | AP (carb$^{R}$) | SC101 | $P_{lacZ\text{-}opt}$ (off-target) | luxAB | $P_{pro1}$ | 434cI-SH2$_{ABL1}$ | - | - | 2A |
| pAB078d8 | AP (carb$^{R}$) | SC101 | $P_{lacZ\text{-}opt}$ (OR1+2) | luxAB | $P_{pro1}$ | 434cI-SH2$_{ABL1}$ | - | - | 2A, S3C |
| pAB078d9 | AP (carb$^{R}$) | SC101 | $P_{lacZ\text{-}opt}$ (OR1+2+3) | luxAB | $P_{pro1}$ | 434cI-SH2$_{ABL1}$ | - | - | 2A |
| pAB082a | CP (spec$^{R}$) | ColEI | $P_{pro1}$ | rpoZ-Cry1Ac-d123 | - | - | - | | S5C |
| pAB082b | CP (spec$^{R}$) | ColEI | $P_{pro1}$ | rpoZ-Cry1Ac-d3 | - | - | - | | S5C |
| pAB082c | CP (spec$^{R}$) | ColEI | $P_{pro1}$ | rpoZ-Cry1Ac-d2A | - | - | - | | S5C |
| pAB082d | CP (spec$^{R}$) | ColEI | $P_{pro1}$ | rpoZ-Cry1Ac-d2B | - | - | - | | S5C |
| pAB082e | CP (spec$^{R}$) | ColEI | $P_{pro1}$ | rpoZ-Cry1Ac-d2C | - | - | - | | S5C |
| pAB082f | CP (spec$^{R}$) | ColEI | $P_{pro1}$ | rpoZ-Cry1Ac-d23A | - | - | - | | S5C |
| pAB082g | CP (spec$^{R}$) | ColEI | $P_{pro1}$ | rpoZ-Cry1Ac-d23B | - | - | - | | S5C |
| pAB082h | CP (spec$^{R}$) | ColEI | $P_{pro1}$ | rpoZ-Cry1Ac-d23C | - | - | - | | S5C |
| pAB085c | AP (carb$^{R}$) | SC101 | $P_{lacZ\text{-}opt}$ (OR1) | luxAB | $P_{tet}$ | 434cI-TnTBR3-FL | - | | S5C |
| pAB085d | AP (carb$^{R}$) | SC101 | $P_{lacZ\text{-}opt}$ (OR1) | luxAB | $P_{tet}$ | 434cI-TnTBR3-F3 | - | | S5C |
| pAB085d6 | AP (carb$^{R}$) | SC101 | $P_{lacZ\text{-}opt}$ (OR1) | luxAB | $P_{pro1}$ | 434cI-TnTBR3-F3 | - | - | 3B |
| pAB085d7 | AP (carb$^{R}$) | SC101 | $P_{lacZ\text{-}opt}$ (OR1) | luxAB | $P_{pro1}$ | 434cI-TnCAD-F3 | - | - | 3C |
| pAB085e | AP (carb$^{R}$) | SC101 | $P_{lacZ\text{-}opt}$ (OR1) | luxAB | $P_{tet}$ | 434cI-TnTBR3-F7 | - | | S5C |
| pAB088c | AP (carb$^{R}$) | SC101 | $P_{lacZ\text{-}opt}$ (OR1) | gIII, luxAB | $P_{pro1}$ | 434cI-TnTBR3-F3 | - | - | 3A |
| pAB088e | AP (carb$^{R}$) | SC101 | $P_{lacZ\text{-}opt}$ (OR1+2) | gIII, luxAB | $P_{pro1}$ | 434cI-TnTBR3-F3 | - | - | 3A |
| pAB088h | AP (carb$^{R}$) | SC101 | $P_{lacZ\text{-}opt}$ (OR1+2) | gIII, luxAB | $P_{pro1}$ | 434cI-TnCAD-F3 | - | - | 3A |
| pAB088i | AP (carb$^{R}$) | SC101 | $P_{lacZ\text{-}opt}$ (OR1) | gIII, luxAB | $P_{pro1}$ | 434cI(RR69)-TnCAD-F3 | - | - | 3A |
| pAB092a | AP (carb$^{R}$) | SC101 | $P_{lacZ\text{-}opt}$ (OR1+2) | gIII, luxAB | $P_{pro1}$ | 434cI-SH2$_{ABL1}$ | - | - | 2B |
| pAB094a | CP (spec$^{R}$) | ColEI | $P_{BAD}$ | rpoZ-HA4 | $P_{C}$ | araC | - | - | S3B |
| pAB094b | CP (spec$^{R}$) | ColEI | $P_{BAD}$ | rpoZ-HA4 (Y35A) | $P_{C}$ | araC | - | - | S3B |
| pAB094c | CP (spec$^{R}$) | ColEI | $P_{BAD}$ | rpoZ-HA4 (R38A) | $P_{C}$ | araC | - | - | S3B |
| pAB094d | CP (spec$^{R}$) | ColEI | $P_{BAD}$ | rpoZ-HA4 (E52A) | $P_{C}$ | araC | - | - | S3B |
| pAB094e | CP (spec$^{R}$) | ColEI | $P_{BAD}$ | rpoZ-HA4 (Y87A) | $P_{C}$ | araC | - | - | S3B |
| pAB107a | AP (carb$^{R}$) | SC101 | $P_{lacZ\text{-}opt}$ (OR1) | gIII, luxAB | $P_{pro1}$ | 434cI | - | - | 2B |
| pJC175e | AP (carb$^{R}$) | SC101 | $P_{PSP}$ | gIII, luxAB | - | - | - | - | 2B |
| SP013 | SP (kan$^{R}$) | M13 f1 | $P_{gIII}$ | rpoZ | - | - | - | | S5D, S5E |
| SP055 | SP (kan$^{R}$) | M13 f1 | $P_{gIII}$ | rpoZ-Cry1Ac | - | - | - | | 3A, S5D, S5E |
| SP096 | SP (none) | M13 f1 | $P_{gIII}$ | rpoZ-HA4 | - | - | - | - | 2B |
| SP097 | SP (none) | M13 f1 | $P_{gIII}$ | rpoZ-HA4 (Y87A) | - | - | - | - | 2B |
| SP098 | SP (none) | M13 f1 | $P_{gIII}$ | rpoZ | - | - | - | - | 2B |
| MP4 | MP (chlor$^{R}$) | CloDF13 | $P_{BAD}$ | dnaQ926, dam, seqA | $P_{C}$ | araC | - | - | 3A |
| MP6 | MP (chlor$^{R}$) | CloDF13 | $P_{BAD}$ | dnaQ926, dam, seqA, emrR, ugi, cda1 | $P_{C}$ | araC | - | - | 3A |
| pMON101647 | EP (chlor$^{R}$) | ColEI | $P_{T7}$ | (none) | - | - | - | - | - |
| pMON101695 | EP (chlor$^{R}$) | P15A | $P_{T7}$ | MBP-TVMV | - | - | - | - | - |
| pMON133051 | EP (chlor$^{R}$) | ColEI | $P_{T7}$ | Cry1Ac | - | - | - | - | - |
| pMON251427 | EP (kan$^{R}$) | ColEI | $P_{T7}$ | TnTBR3-FL | - | - | - | - | - |
| IS0008 | EP (kan$^{R}$) | ColEI | $P_{T7}$ | TnCAD-FL | - | - | - | - | - |
| pMON262346 | EP (chlor$^{R}$) | ColEI | $P_{T7}$ | Cry1Ac | - | - | - | - | - |

Each plasmid is defined by the plasmid class, antibiotic resistance, origin of replication, and promoter/gene combinations describing the relevant open reading frames carried by the plasmid. Relevant figures where these materials were used are given in each case.

LETTER

doi:10.1038/nature17417

# Resolved atomic lines reveal outflows in two ultraluminous X-ray sources

Ciro Pinto[1], Matthew J. Middleton[1] & Andrew C. Fabian[1]

**Ultraluminous X-ray sources are extragalactic, off-nucleus, point sources in galaxies, and have X-ray luminosities in excess of $3 \times 10^{39}$ ergs per second. They are thought to be powered by accretion onto a compact object. Possible explanations include accretion onto neutron stars with strong magnetic fields[1], onto stellar-mass black holes (of up to 20 solar masses) at or in excess of the classical Eddington limit[2–4], or onto intermediate-mass black holes ($10^3$–$10^5$ solar masses)[5]. The lack of sufficient energy resolution in previous analyses has prevented an unambiguous identification of any emission or absorption lines in the X-ray band, thereby precluding a detailed analysis of the accretion flow[6–8]. Here we report the presence of X-ray emission lines arising from highly ionized iron, oxygen and neon with a cumulative significance in excess of five standard deviations, together with blueshifted (about 0.2 times light velocity) absorption lines of similar significance, in the high-resolution X-ray spectra of the ultraluminous X-ray sources NGC 1313 X-1 and NGC 5408 X-1. The blueshifted absorption lines must occur in a fast-outflowing gas, whereas the emission lines originate in slow-moving gas around the source. We conclude that the compact object in each source is surrounded by powerful winds with an outflow velocity of about 0.2 times that of light, as predicted by models of accreting supermassive black holes and hyper-accreting stellar-mass black holes[9,10].**

NGC 1313 X-1, NGC 5408 X-1 and NGC 6946 X-1 are three ultraluminous X-ray sources (ULXs), with X-ray luminosities up to $\sim 10^{40}$ erg s$^{-1}$. All three sources have been observed in spectral 'states' in which a large proportion of the flux emerges at energies below 2 keV (refs 6, 11) and in these states they show strong spectral deviations (0.6–1.2 keV) from the underlying continuum in charge-coupled device (CCD) spectra[8,12]. Owing to their relative proximity to Earth (distances $D < 7$ Mpc) and brightness, data quality is high, making them ideal targets for understanding the spectral residuals through the high energy-resolution reflection grating spectrometer (RGS) on board ESA's XMM-Newton observatory.

NGC 1313 has the lowest star-formation rate of the three galaxies, and its X-1 is well isolated, being located several arcminutes from the other X-ray bright sources in the galaxy (X-2 and SN 1978K, see Extended Data Fig. 1). Archival Chandra observations (with sub-arcsecond spatial resolution) show it to be confined within a region of 6 arcsec radius (<116 pc at a distance of 3.95 Mpc, see Extended Data Fig. 2). This is not the case for NGC 5408 X-1, whose spectra are affected by a nearby X-ray bright source (see Extended Data Fig. 3a). NGC 6946 has a much high star-formation rate, shorter observations and the ULX X-1 shows a weaker X-ray continuum (see Extended Data Table 1). NGC 1313 X-1 therefore allows for a comparatively 'clean', high energy-resolution study of the features seen in CCD spectra[12].

XMM-Newton has observed NGC 1313 several times in the last 15 years with three observations centred on X-1 each lasting ~100 ks, providing independent, well exposed, high-resolution RGS spectra (see Extended Data Table 1). We extract the RGS spectra (see Methods for details) and identify strong, rest-frame emission lines from a mixture of elements at varying degrees of ionization, including Ne X (wavelength 12.1 Å), O VIII (19.0 Å) and O VII (21.6 Å) resonance lines (see Fig. 1 and Extended Data Figs 4 and 5) with the blue side of Ne X partly absorbed. We also find evidence for Fe XVII resonance (15.0–15.3 Å) and forbidden (17.1 Å) lines. We apply a series of physically consistent models for the lines to both the EPIC (European Photon Imaging PN Camera) and RGS data simultaneously (see Methods; http://www.sron.nl/spex). The deviations from the blackbody and power-law continuum model previously seen in bright ULXs[8,12] are now resolved by the inclusion of the RGS data into a complex of emission and absorption lines (see Fig. 1). The emission lines are highly significant (at $>3\sigma$ each for Ne X, Fe XVII and O VIII and $>5\sigma$ in total) and can be well modelled with a rest-frame, collisional ionization equilibrium (CIE) (gas), which includes an underlying weak bremsstrahlung continuum at an average temperature of 0.8 keV ($\sim 10^7$ K; see Fig. 1 and Extended Data Table 2). The absorption lines (significant at $5\sigma$ in total) can be well modelled with two-phase, low ionization absorbing gas in photoionization equilibrium applied to the continuum. One absorber is consistent with being at rest, while the other requires a high outflow velocity of around 0.2$c$. More details on the spectral modelling (and abundance ratios) are reported in Methods. The inclusion of a third, velocity-broadened absorber[8] significantly improves the fit relative to the continuum model. This model requires moderately relativistic velocities ($\sim 0.25c$), a high column density ($N_H \approx 1 \times 10^{24}$ cm$^{-2}$) and high ionization parameter ($\xi \approx 3 \times 10^4$ erg cm s$^{-1}$). It is described by the blue line in Fig. 1. The firm detection and identification of rest-frame emission and blueshifted absorption lines open up new and powerful means to understand ULXs.

The three individual observations of NGC 1313 X-1 show evidence for line variability (see, for example, Ne X in Extended Data Fig. 5). Absorption is detected in the first two observations, while the emission lines are stronger in observation 3 where their flux is twice that seen in observation 1 (see Methods). Emission lines are weaker in observation 2 and show a decrease in the ionization parameter. We do not detect significant absorption in observation 3. In Fig. 2 we show the ratios of the RGS spectra between the individual exposures, which confirm the variability of both absorption and emission lines. We do not find a significant trend in the strength of the features with the spectral hardness of the source, most probably because the spectral states of the three observations are very similar[12].

Spectral fits performed using only the RGS data (that is, excluding the PN data) from the three 100 ks exposures confirmed the detection of the emission and absorption components (see Methods). In Fig. 3 we show the significance obtained adopting 500 km s$^{-1}$ and 10,000 km s$^{-1}$ line widths; negative values indicate absorption lines. Each emission line is detected individually at $3\sigma$, confirming the $>5\sigma$ detection obtained with the CIE emission model (which treats all of the lines consistently). Ne IX and Ne X blueshifted absorption is also individually detected between $3\sigma$ and $5\sigma$, in agreement with the photoionization code.

[1]Institute of Astronomy, Cambridge University, Madingley Road, Cambridge CB3 0HA, UK.

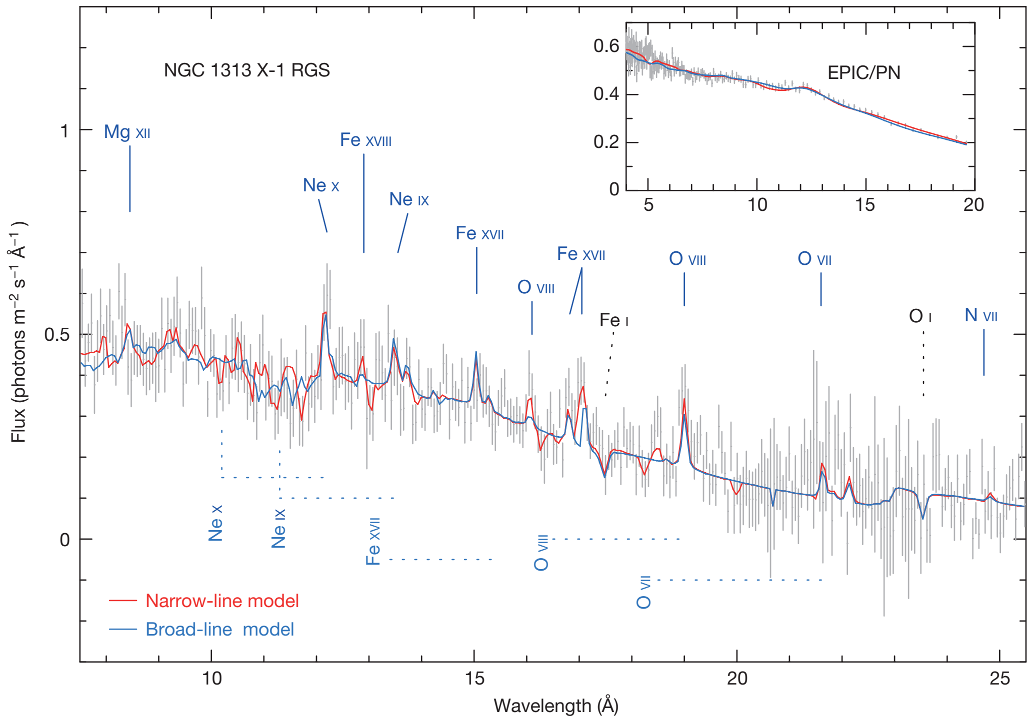



**Figure 1 | Simultaneous spectral fits to the stacked XMM-Newton RGS and EPIC/PN spectra of NGC 1313 X-1.** Main panel, the RGS stacked spectrum; inset, the PN stacked spectrum (same variables on axes). The rest-frame wavelengths of the most relevant transitions and some blueshifted lines are labelled (the dashed lines show the velocity shift). An isothermal emission model of gas in collisional ionization equilibrium describes most emission lines at rest. The absorption lines are reproduced with multi-phase models for gas in photoionization equilibrium. Red line, model consisting of rest-frame absorption and emission and a relativistically outflowing ($v = 0.2c$) photoionized absorber. Blue line, model that includes an additional broadened absorber ($v = 0.25c$). Error bars, $\pm 1\sigma$.

The emission lines show comparable fluxes of $2.5 \times 10^{-6}$ photons s$^{-1}$ cm$^{-2}$ and equivalent widths of 15–30 mÅ, while the absorption lines have equivalent widths from 15 mÅ up to 250 mÅ. Lines associated with warm absorbers around active galactic nuclei typically have similar equivalent widths[13], but different dominant species: the O VII and Ne IX triplets and the common Fe XVII unresolved transition array (15–17 Å)[13]. Galactic X-ray binaries and microquasars also show comparable equivalent widths[14,15,16] and similar ionic species, for example, Ne X and O VIII, but intercombination transitions may play an important role[17].

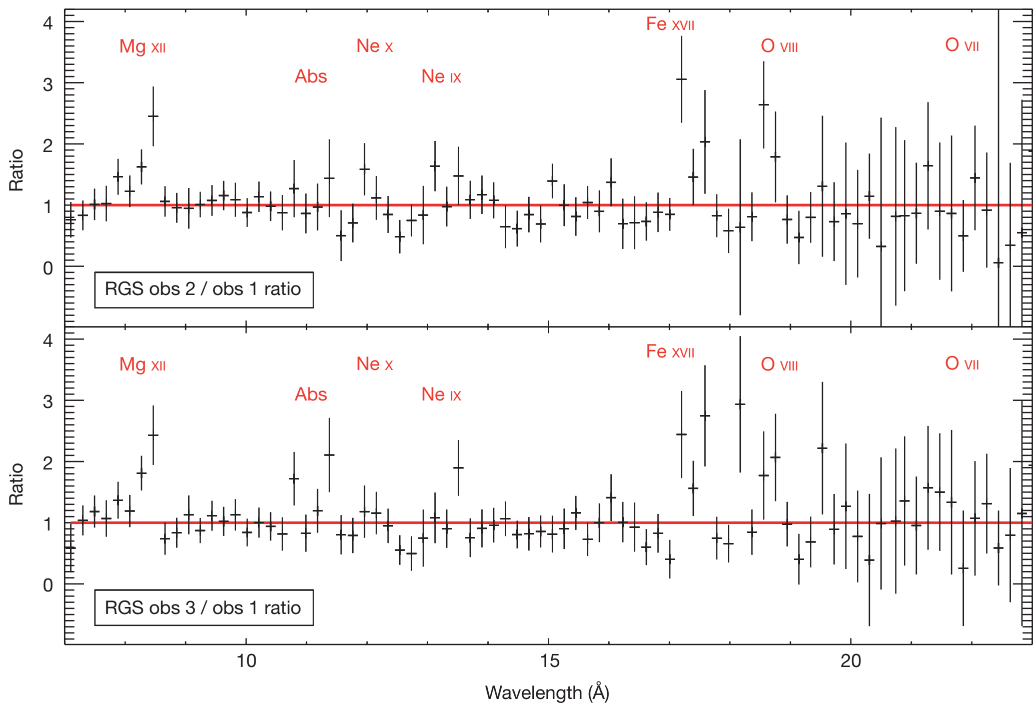



**Figure 2 | Ratios between the individual RGS spectra of NGC 1313 X-1.** The RGS spectra were normalized by the spectral continuum and divided by that of observation 1 (obs 1). The absorption features ('Abs', 10.7–11.4 Å) change in observation 3 (obs 3). Rest-frame emission features also exhibit variability (see also Extended Data Fig. 5). Error bars, $\pm 1\sigma$.

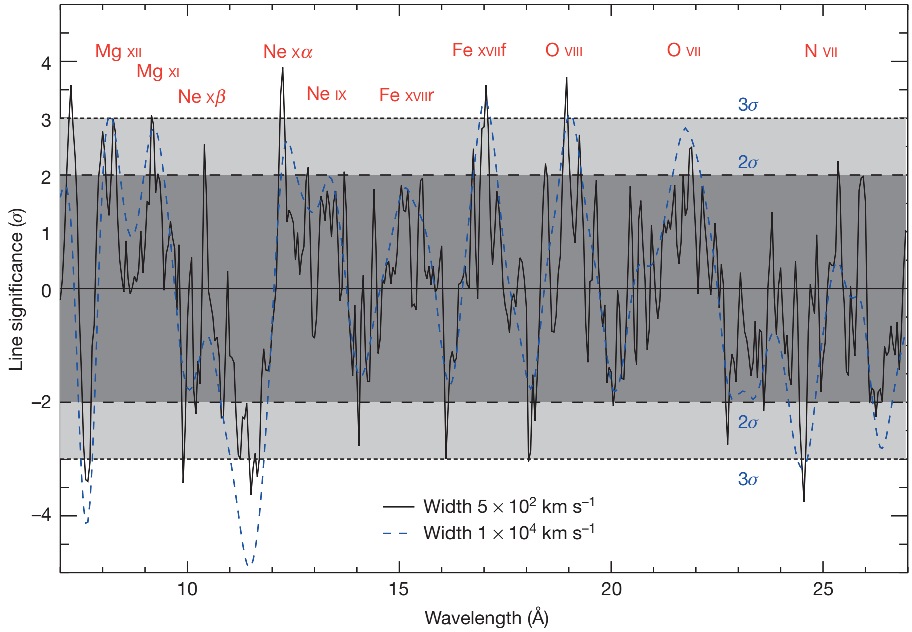



**Figure 3 | Significance of the features in the NGC 1313 X-1 RGS stacked spectrum.** Shown is the line significance obtained by Gaussian fitting over the 7–27 Å wavelength range with increments of 0.05 Å and negative values indicating absorption lines. The solid and dashed lines indicate the line significance obtained with 500 km s$^{-1}$ and 10,000 km s$^{-1}$ widths, respectively. Ne X, Fe XVII, and O VIII emission lines are individually detected at 3$\sigma$, and combined provide an 8$\sigma$ detection. Ne X blueshifted absorption is clearly detected up to 5$\sigma$, showing widths larger than the emission lines.

The emission lines in NGC 1313 X-1 differ from those typically seen in active galactic nuclei and some X-ray binaries, but they are very similar to those produced by the accretion disk of the X-ray binary 4U 1626-67[17], which may suggest an origin in a wind launched by a compact disk. The features seen in NGC 1313 X-1 are more difficult to detect than those seen in X-ray binaries because of the several orders of magnitude difference in distance (from a kiloparsec scale for the binaries in our Galaxy to the megaparsec scale for NGC 1313 X-1).

While the emission lines are seen at their rest-frame energies, the blueshifted absorption lines confirm the presence of an outflow, that is, photoionized gas within a wind[8,12]. The high ionization parameter and outflow velocity suggests an accretion-disk-wind origin similar to that of Galactic black hole binaries[14], but with far larger velocities and therefore energetics (see Methods for possible interpretations). The fact that we detect both emission and absorption lines—the latter being line-of-sight dependent—requires that NGC 1313 X-1 is being seen at a moderate inclination angle (assuming an equatorial wind)[12],

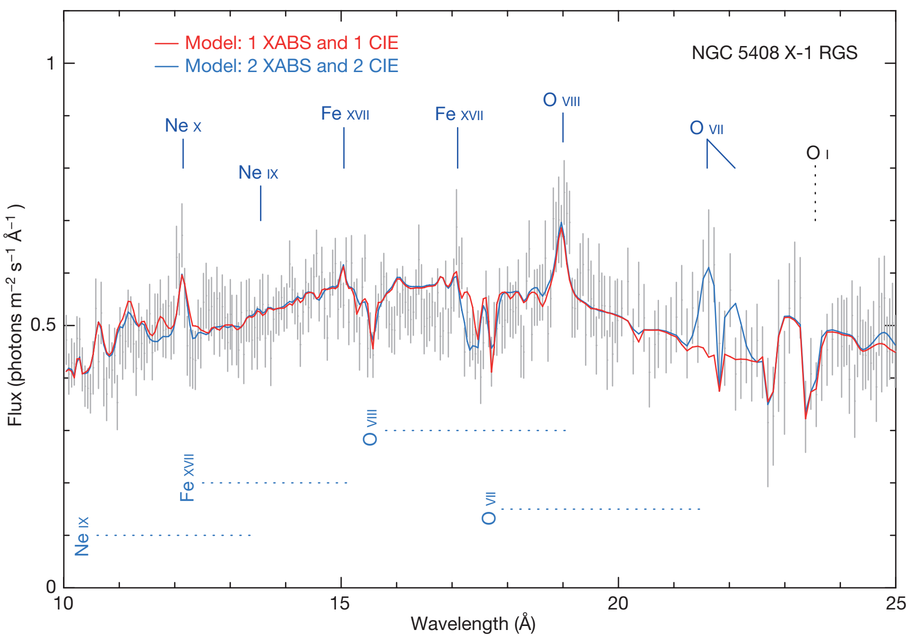



**Figure 4 | Best fit to the stacked XMM-Newton RGS spectrum of NGC 5408 X-1.** Line labels are same as in Fig. 1. An isothermal emission model of gas in collisional ionization equilibrium describes most emission lines at rest. The absorption lines can be reproduced with gas in photoionization equilibrium. The red line is a model consisting of a single, relativistically outflowing ($v = 0.22c$), photoionized absorber. The blue line is a model that includes two absorbers ($v_1 = 0.10c$ and $v_2 = 0.22c$). The absorption lines have widths of 500 ± 300 km s$^{-1}$, while the emission lines are broader with $\sigma_v = 2{,}000 \pm 500$ km s$^{-1}$. Error bars, ±1$\sigma$.

with the difference between the widths and the Doppler shifts of the emission and absorption lines suggesting different spatial locations.

If collisional equilibrium applies, the most obvious explanation for the emission lines is either shock or collisional heating in the outflow with a range of velocities or between the outflow and the wind of the stellar companion as commonly seen in colliding-wind binaries[18,19]. Such lines should also be present in other ULXs, but absorption features may be harder to detect if there is source confusion or if the spectral 'state' is hard[12].

NGC 5408 X-1 also shows sharp emission features similar to those detected in NGC 1313 X-1, including the Ne X and the O VIII resonance lines and narrow absorption features at 12.2 Å, 15.5 Å and 17–18 Å (see Fig. 4, Extended Data Fig. 6, and Methods for details). The strongest emission lines are individually detected at $5\sigma$, while the absorption features have lower significance (about $4\sigma$ in total). Most features can be described by an isothermal emission model of gas in collisional ionization equilibrium (temperature $T \approx 3$ keV) and a relativistically outflowing (velocity $v = 0.22c$) photoionized gas model (see red line in Fig. 4). The main absorbers in NGC 5408 X-1 and NGC 1313 X-1 have comparable outflow velocities ($v \approx 0.2c$), suggesting that they could have the same origin although the wind may be more structured in NGC 5408 X-1 (see Methods).

NGC 6946 X-1 exhibits the O VIII (19.0 Å) and Ne IX (13.45 Å) emission lines and a feature at $\sim$11.0 Å that could be attributed to either higher-ionization Fe XXII–XXIII emission lines or to an absorption edge (see Extended Data Fig. 4). Its very low continuum prevents the detection of any absorption lines.

The emission lines that appear in the two ultraluminous X-ray sources NGC 1313 X-1 and NGC 5408 X-1 are probably associated with collisional shock heating between the circumsystem gas and the outflowing wind that we have now identified in the form of absorption lines. This result suggests that the accretion flow in some ULXs can be associated with powerful winds that leave their imprint in emission and absorption lines and are able to produce the common residuals in the high-quality CCD-resolution spectra of the most bright, well studied ULXs, for example, NGC 1313 X-1, Ho IX X-1, Ho II X-1, NGC 55 X-1, NGC 5204 X-1, NGC 5408 X-1 and NGC 6946 X-1[12].

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 29 July 2015; accepted 4 February 2016.**
**Published online 27 April 2016.**

1. Bachetti, M. *et al.* An ultraluminous X-ray source powered by an accreting neutron star. *Nature* **514,** 202–204 (2014).
2. Shakura, N. I. & Sunyaev, R. A. Black holes in binary systems: observational appearance. *Astron. Astrophys.* **24,** 337–355 (1973).
3. Poutanen, J. *et al.* Supercritically accreting stellar mass black holes as ultraluminous X-ray sources. *Mon. Not. R. Astron. Soc.* **377,** 1187–1194 (2007).
4. King, A. R. *et al.* Ultraluminous X-ray sources in external galaxies. *Astrophys. J.* **552,** L109–L112 (2001).
5. Pasham, D. R., Strohmayer, T. E. & Mushotzky, R. F. A 400 solar mass black hole in the ultraluminous X-ray source M82 X-1 accreting close to its Eddington limit. *Nature* **513,** 74–76 (2014).
6. Stobbart, A.-M., Roberts, T. P. & Wilms, J. XMM-Newton observations of the brightest ultraluminous X-ray sources. *Mon. Not. R. Astron. Soc.* **368,** 397–413 (2006).
7. Bachetti, M. *et al.* The ultraluminous X-ray sources NGC 1313 X-1 and X-2: a broadband study with NuSTAR and XMM-Newton. *Astrophys. J.* **778,** 163–173 (2013).
8. Middleton, M. J., Walton, D. J., Roberts, T. P. & Heil, L. Broad absorption features in wind-dominated ultraluminous X-ray sources? *Mon. Not. R. Astron. Soc.* **438,** L51–L55 (2014).
9. King, A. & Pounds, K. Powerful outflows and feedback from active galactic nuclei. *Annu. Rev. Astron. Astrophys.* **53,** 115–154 (2015).
10. King, A. & Muldrew, S. I. Black hole winds II: hyper-Eddington winds and feedback. *Mon. Not. R. Astron. Soc.* **455,** 1211–1217 (2016).
11. Gladstone, J. C., Roberts, T. P. & Done, C. The ultraluminous state. *Mon. Not. R. Astron. Soc.* **397,** 1836–1851 (2009).
12. Middleton, M. J. *et al.* Diagnosing the accretion flow in ULXs using soft X-ray atomic features. *Mon. Not. R. Astron. Soc.* **454,** 3134–3142 (2015).
13. Kaspi, S. *et al.* The ionized gas and nuclear environment in NGC 3783. I. Time-averaged 900 kilosecond Chandra grating spectroscopy. *Astrophys. J.* **574,** 643–662 (2002).
14. Ponti, G. *et al.* Ubiquitous equatorial accretion disc winds in black hole soft states. *Mon. Not. R. Astron. Soc.* **422,** L11–L15 (2012).
15. Miller, J. M. *et al.* The accretion disk wind in the black hole GRO J1655–40. *Astrophys. J.* **680,** 1359–1377 (2008).
16. Marshall, H. L., Canizares, C. R. & Schulz, N. S. The high-resolution X-ray spectrum of SS 433 using the Chandra HETGS. *Astrophys. J.* **564,** 941–952 (2002).
17. Schulz, N. S. *et al.* Double-peaked X-ray lines from the oxygen/neon-rich accretion disk in 4U 1626–67. *Astrophys. J.* **563,** 941–949 (2001).
18. Cooke, B. A., Fabian, A. C. & Pringle, J. E. Upper limits to X-ray emission from colliding stellar winds. *Nature* **273,** 645–646 (1978).
19. Oskinova, L. M. Evolution of X-ray emission from young massive star clusters. *Mon. Not. R. Astron. Soc.* **361,** 679–694 (2005).

**Acknowledgements** We thank A. Lohfink for help with the use of photoionization emission codes. A.C.F. acknowledges support from the European Research Council through Advanced Grant on Feedback 340492. M.J.M. appreciates support from an STFC advanced fellowship. This work is based on observations with XMM-Newton, an ESA science mission with instruments and contributions directly funded by ESA Member States and NASA. This research has also made use of data obtained from NASA's Chandra satellite. All codes used are publicly available.

**Author Contributions** C.P. wrote the manuscript with comments from all authors, and analysed the XMM-Newton data. Both M.J.M. and A.C.F. made substantial contributions to the overall science case and manuscript.

**Author Information** Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to C.P. (cpinto@ast.cam.ac.uk).

## METHODS

**Data reduction.** The XMM-Newton satellite is equipped with two types of X-ray detectors: the CCD-type European Photon Imaging Cameras (EPICs)[20,21] and the Reflection Grating Spectrometers (RGSs)[22]. The EPICs are MOS and PN. The RGS camera consists of two similar detectors, which have high effective area and high spectral resolution between 6 Å and 38 Å.

All the observations of the sources have been reduced with the XMM-Newton Science Analysis System (SAS) v13.5.0 (http://www.cosmos.esa.int/web/xmm-newton/sas). We correct for contamination from soft-proton flares following the XMM-SAS standard procedures. For each source and exposure, we extracted the first-order RGS spectra in a cross-dispersion region of 1 arcmin width, centred on the emission peak. We have extracted background spectra by selecting photons beyond the 98% of source point-spread-function. The background spectra were comparable to those from blank field observations. We extracted the MOS and PN images in the RGS (0.35–1.8 keV) energy band and stacked them all with the *emosaic* SAS task (see Extended Data Figs 1 and 3). We also extracted EPIC MOS and PN spectra from within a circular region of 1 arcmin diameter centred on the emission peak. The background spectra were extracted from within a 1 arcmin circle in a nearby region on the same chip, but away from bright sources and the readout direction. As the EPIC/PN spectra contain the majority of the counts and the residuals have been shown to not be instrumental in origin[23], we discard the EPIC/MOS spectra from our analysis. The total clean exposure times are quoted in Extended Data Table 1.

**Sample size.** No statistical methods were used to predetermine sample size.

**EPIC + RGS spectral modelling.** We fit the EPIC/PN and RGS spectra (with the SPEX package; http://www.sron.nl/spex) simultaneously to constrain both the broad-band continuum and describe the atomic features. Importantly, we fit across individual spectra in each observation rather than stacking the data in order to avoid any spurious features resulting from different pointing and background subtraction, which differ between RGS 1 and 2. We bin both the RGS and PN spectra in channels equal to 1/3 of the PSF, and use C-statistics, because it provides the optimal spectral binning and avoids over-sampling.

The phenomenological continuum model we apply to the data is a combination of soft blackbody with temperature $T \approx (2.5\text{–}3.0) \times 10^6$ K and power-law emission components with photon index $\Gamma \approx 1.9$ extending to high energies, both absorbed by neutral gas with a best-fit hydrogen column density $N_{\rm H} = (1.8 \pm 0.1) \times 10^{21}\,{\rm cm^{-2}}$, which includes any intrinsic and Galactic absorption ($N_{\rm H}^{\rm Gal} = 4 \times 10^{20}\,{\rm cm^{-2}}$)[24] adopting solar abundances[25]. The continuum of the three NGC 1313 X-1 observations shows little evidence for variability above 1 keV while the soft X-ray band is variable. All the model parameters for the RGS and EPIC spectra of the same observation are tied with each other while the continuum parameters (normalization and slope of the power law, normalization and temperature of the blackbody) are uncoupled between different observations. We detect a complex of emission and absorption lines (see Fig. 1). Under the first-order assumption that the lines are unchanging between observations, we also tie the parameters of the absorption and emission-line models for the different observations in order to increase the statistics.

The emission lines resolved by the RGS can be well modelled with a rest-frame, collisionally ionized gas (CIE model in SPEX), which is detected up to $8\sigma$ (see Fig. 1, red line, and Extended Data Table 2). The absorption-like features can be modelled with a two-phase absorbing gas in photoionization equilibrium. This can be described by a combination of two XABS models in SPEX. The $\Delta C_{\rm stat}$ and the equivalent $\Delta\chi^2$ provided by each component, which indicates their improvement to the fit, are reported in Extended Data Table 2. The ionization parameters $\xi$ were tied between the different XABS models because a preliminary fit showed that, if they are left free to vary, they agree within error. One component is consistent with being at rest, while the other requires a high outflow velocity of $\sim 0.2c$ ($v_2 = -65{,}000 \pm 10{,}500\,{\rm km\,s^{-1}}$). The column densities of the two absorbers are $N_{\rm H,1} = (1.5 \pm 0.3) \times 10^{22}\,{\rm cm^{-2}}$ and $N_{\rm H,2} = (5 \pm 1) \times 10^{21}\,{\rm cm^{-2}}$, respectively. At this stage, the velocity broadening was tied between the emission and the absorption components with the best fit giving a $1\sigma$ upper limit of about $500\,{\rm km\,s^{-1}}$. When we untie the velocity broadening between these components, we obtain $v_{\sigma,\rm CIE} = 1{,}000 \pm 500\,{\rm km\,s^{-1}}$ and $v_{\sigma,\rm XABS} < 20\,{\rm km\,s^{-1}}$, respectively. The ionization parameter of the absorbers is rather low ($\xi = 200 \pm 100\,{\rm erg\,cm\,s^{-1}}$).

It is not possible to determine the absolute metal abundances from the RGS lines, as these require comparison to hydrogen lines which are absent in X-ray spectra. In the fits above, the elemental abundances were therefore tied between the emitting and absorbing gas components at solar metallicity for iron[25] with free ratios for the abundances of oxygen, neon, and magnesium (because Fe XVII and Fe XVIII produce detectable lines: see Fig. 1). On average, the emitting and absorbing plasmas exhibit abundance ratios as follows: O/Fe = 1.0 ± 0.2, Ne/Fe = 1.8 ± 0.4, and Mg/Fe = 2.1 ± 0.5. This suggests a small over-abundance of $\alpha$ elements (with respect to iron), which indicates a small amount of supernovae core-collapse (SN cc) enrichment with respect to the solar ratio of SN Ia to SN[25,26]. Most ULXs—including the three studied in this work—are found in spiral/interacting/star-forming galaxies where a recent contribution from SN cc could be expected[27].

To the model described above we add a third photoionized absorber to search for high velocity broadening as invoked in a previous study using the CCD spectra[8,12]. The $v_\sigma$ was therefore untied between all components. A highly significant improvement to the continuum-only model ($\Delta\chi^2$/d.o.f. = 150/4, see Model 2 in Extended Data Table 2) was also obtained with a highly ionized, outflowing (at moderately relativistic velocities $\sim 0.25c$), and optically thick absorber (see blue line in Fig. 1). The large velocity broadening ($\sim 0.1c$) in this latter model may explain why these features appear weak compared to the emission lines, however, an alternative reason may be variability between observations (indeed a trend in the strength of the residuals in the CCD spectra with spectral hardness was recently discovered[12]). There is some degree of degeneracy in the absorption models. The inclusion of the broadened XABS 3 component strongly decreases the significance of the other two components XABS 1–2. Longer exposures are needed to better characterize the outflow.

In principle, the emission lines could also be produced by photoionized gas further away from the X-ray source. Indeed, the Fe XVII 17 Å forbidden (f) line is much stronger than the 15 Å resonance (r) line which would suggest either photoionization or resonant absorption. As SPEX does not provide a model for line-emitting gas in photoionization equilibrium, we used the *photemis* model in XSPEC (http://heasarc.nasa.gov/docs/software/xspec/) to create a grid of photoionization emission models with log $\xi$ from 1.0 to 4.0 with a 0.25 step size. The best fit still supports an Ne/Fe ≥ 1 abundance ratio. It is difficult to distinguish between photoionization and collisional ionization models as the results are comparable; deeper observations are necessary. Alternatively, the emission features may originate from recombination of highly ionized gas within the wind or in a distant region, which is expected if photoionization occurs. However, a recombination model does not describe the lines satisfactorily. A shock between the outflow and the low density material in the surrounding nebula is also ruled out owing to the substantial X-ray brightness and the size of the X-ray source (see Extended Data Fig. 2) which is far more spatially compact (<116 pc) than the surrounding nebula (240–800 pc)[28].

**RGS-only spectral modelling.** We performed RGS-only spectral fits to check the line detection. We removed the EPIC-PN data and froze the continuum parameters to the values obtained with the simultaneous EPIC-RGS fit. We used Model 1 (with two narrow absorbers) and confirmed the need for both rest-frame emission and absorption lines. Despite the larger count rate, EPIC-PN does not change the detection significance enormously owing to its poorer spectral resolution with respect to RGS in the soft X-ray band.

The individual RGS spectra for each observation of NGC 1313 X-1 show changes in absorption line strengths (see Extended Data Fig. 5). In order to study the variability of the features, we have fitted the RGS spectra for the individual exposures with a simple model consisting of one CIE line-emitting component and one XABS absorber. Absorption is detected in the first two observations with consistent parameters ($N_{\rm H} = (2.0 \pm 0.4) \times 10^{22}\,{\rm cm^{-2}}$, $\xi = 200 \pm 70\,{\rm erg\,cm\,s^{-1}}$, $v = -57{,}000 \pm 500\,{\rm km\,s^{-1}} \approx 0.2c$, with significance $>3\sigma$ in total for each observation), while the emission lines are stronger in observation 3 where their flux is twice than that seen in observation 1, but the temperature is consistent at $kT = 1.10 \pm 0.15$ keV ($\sim 1.3 \times 10^7$ K). Emission lines are weaker in observation 2 and show a decrease in the ionization parameter where the Fe XVII and O VII lines (from cooler gas) are stronger than the Ne X and O VIII lines. We do not detect significant absorption in observation 3.

**RGS line significance.** We have also confirmed the detection of each emission/absorption line by fitting the RGS spectra adopting the EPIC-RGS continuum and including a Gaussian spanning the 7–27 Å wavelength range in increments of 0.05 Å. We assumed a grid of linewidths from $500\,{\rm km\,s^{-1}}$ (~RGS resolution) to $75{,}000\,{\rm km\,s^{-1}}$ ($0.25c$). In Fig. 3 we show the significance obtained adopting $500\,{\rm km\,s^{-1}}$ and $10{,}000\,{\rm km\,s^{-1}}$ line widths, confirming the lines detected with the CIE emission model. Line broadening does not have a major effect on the detection. The absorption lines have a lower significance because velocity shift is an additional parameter. The strongest feature at 11.5 Å (identified as Ne IX blueshifted absorption) has a chi-squared $P$-value of $6 \times 10^{-7}$ (that is, $5\sigma$), whether we consider it as a sum of 2 strong narrow lines or a single broad line. However, if we take into account all the trials due to the spectral resolution bins and widths, we obtain a probability of $\sim 3 \times 10^{-6}$, which is above $4\sigma$. If we also include the other strong blueshifted lines that are found at exactly the same velocity, for example, O VIII at 16.0 Å and O VII at 18.0 Å, then we obtain a total significance above $5\sigma$.

In order to further check the robustness of our results, we adopt different Ne, Fe, and O abundances for the neutral absorbing gas. The neutral gas of NGC 1313

provides the bulk of the $N_H$ and may have non-solar abundances; this in principle could affect the detection of features in the soft X-ray spectra[8,29]. We have therefore re-fitted the RGS and, afterwards, the EPIC-RGS spectra, simultaneously, with interstellar abundances ranging from $0.1 \times$ to $2.0 \times$ solar. No significant difference was found and the detection level of the lines is unchanged; this was expected for several reasons: the strongest features imprinted by neutral gas are expected between 22.7 Å and 23.5 Å (oxygen K edge and $1s$–$2p$ line)[30] and the lines in the ULX spectra avoid the edges. In addition, as we anticipated, the lines are narrow and their detection is not affected by the continuum-like hydrogen absorption.

We have stacked the first order RGS 1-2 spectra from the individual exposures of the same source for plotting purposes only (the stacked spectrum has a much higher S/N ratio and simplifies the recognition of the lines). We have used the following advanced method to combine fluxed spectra[31] (that is, spectra in flux units). We first created individual fluxed spectra using the SAS task *rgsfluxer* and then averaged them with the SPEX tool *rgs_fluxcombine (option 1)* for RGS 1 and RGS 2, separately. We then ran again the *rgs_fluxcombine (option 2)* to combine the stacked RGS 1 and 2 fluxed spectra into a final RGS spectrum for each ULX. Finally, we used the SPEX task *rgs_fmat* to produce the response matrix for the stacked fluxed spectrum (see the SPEX manual; http://www.sron.nl/spex). The stacked RGS spectra of the three ULXs are shown in Figs 1–4, and in Extended Data Fig. 4.

**Constraints on the energetics of the wind, and the black hole mass.** Here we try to place some constraints on the location of the wind seen in NGC 1313 X-1 as well as the black hole mass, using as a template the parameters estimated for the extreme absorber, XABS 3 (see Extended Data Table 2). The ionization parameter is defined as $\xi = L_{\rm ion}/(n_{\rm H}R^2) = L_{\rm ion}/(n_{\rm H}\Delta R \times R) \times (\Delta R/R)$, where $L_{\rm ion}$ is the 1–1,000 Ry ionizing luminosity of the source, and $\Delta R$, $R$, and $n_{\rm H}$ are the thickness, size, and number density of the absorbing region. This leads to $R = L_{\rm ion}/(N_{\rm H}\xi) \times (\Delta R/R)$, where $N_{\rm H}$ is the column density. Since $\Delta R < R$, then $R < L_{\rm ion}/(N_{\rm H}\xi) = 3 \times 10^{11}$ cm, but $R$ must also be larger than the Schwarzschild radius $R_{\rm S} = 2GM/c^2$. Assuming the escape velocity to be equal to the wind speed ($0.2c$), or in other words that the wind comes from a region where its speed equals the escape velocity, we obtain $R/R_{\rm S} \leq 25$, which provides an upper limit on the black hole mass of 40,000 solar masses (for a region with thickness comparable to its size). A black hole with a stellar mass, that is, up to 100 solar masses, would imply a very thin region ($\Delta R << R$, see Extended Data Fig. 7). Throughout this calculation we adopted unity covering fraction.

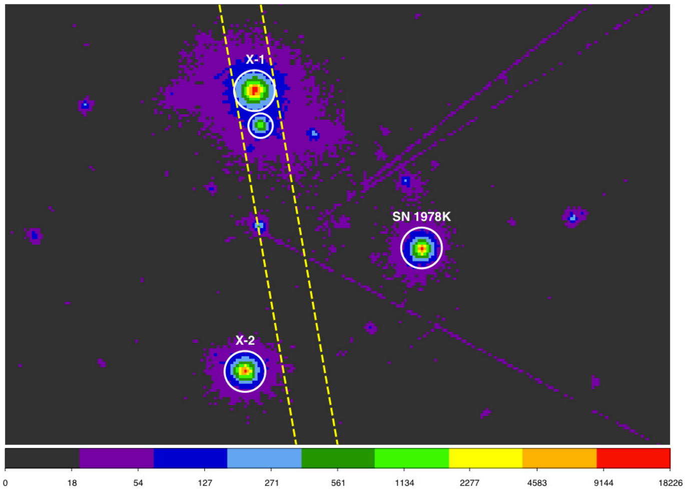
It is interesting to compare the wind power to the source luminosity[32]. The outflow rate can be written as $\dot{M} = 4\pi R^2 \rho v \Omega$, which gives a wind power $P_{\rm w} = 0.5\dot{M}v^2 = 2\pi R^2 m_{\rm p} n_{\rm H} v^3 \Omega$, where $m_{\rm p}$ is the proton mass and $\Omega$ the solid angle. Since $\xi = L/n_{\rm H}R^2$, we get $P_{\rm w} = 2\pi L m_{\rm p} v^3 \Omega/\xi$, which for component XABS 3, provides $P_{\rm w}/L \approx 100\Omega$. This would imply a highly super-Eddington accretion rate, but could be regarded as an upper limit because a smaller outflow rate and kinetic power are obtained if either the covering fraction is lower than unity or the duty cycle is shorter[12]. On the other hand, the wind speed that we measure is a lower limit because it is only maximal for sightlines into the direction of outflow. With the present data characterized by only a few unevenly sampled observations, we cannot accurately measure these parameters.

**NGC 5408 X-1 spectral modelling.** An X-ray image of the ultraluminous X-ray source in NGC 5408 is shown in Extended Data Fig. 3 along with another bright X-ray source (X-2, hereafter) which is covered by the RGS slit. In order to accurately estimate the RGS spectral continuum, we need to estimate the contribution from each source. We have therefore extracted the EPIC-PN spectra in two circular regions of one arcmin centred on the two sources. The background was chosen from a source-free circular region on the same chip and away from the read-out direction. The EPIC spectrum of NGC 5408 X-1 was modelled with a soft blackbody ($kT = 0.14$ keV $\approx 1.6 \times 10^6$ K) and a power law ($\Gamma \approx 2.6$). The X-2 EPIC spectrum is very well modelled by a single power-law component with $\Gamma = 1.99 \pm 0.01$ and neutral column density of $(5.6 \pm 0.2) \times 10^{20}$ cm$^{-2}$, consistent with the H I maps[24]. No features or residuals are detected in the X-2 EPIC spectrum; this is most probably a background AGN.

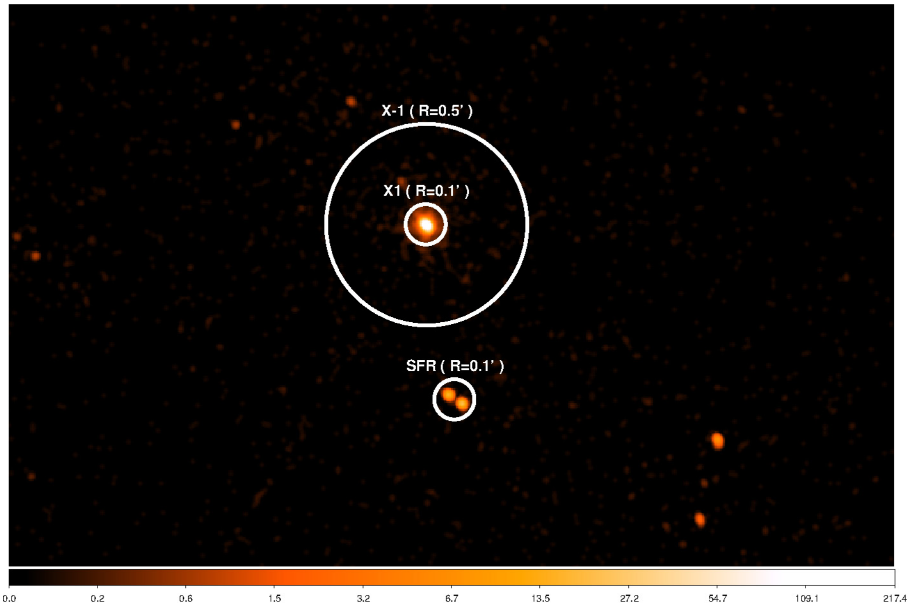
We have built up a spectral model comprising the continuum from both X-1 and X-2 EPIC spectra and applied it to the RGS stacked spectrum of NGC 5408 X-1. We searched for residual emission and absorption features as we did for NGC 1313 X-1 (that is, using a Gaussian line stepping in wavelength). In Extended Data Fig. 6 we show the line significance obtained with 500 km s$^{-1}$ and 10,000 km s$^{-1}$ widths; the results do not strongly depend on the linewidth. The rest-frame wavelengths of some relevant transitions are labelled. Ne X and O VIII emission lines are detected each at $4\sigma$. Blueshifted absorption is also clearly detected with the strongest features detected at $3\sigma$ each (taking into account the number of velocity bins), which provides a detection $>4\sigma$ in total for blueshifted absorption at a velocity shift of about 66,000 km s$^{-1}$.

We proceed to test physical models for the features, first with an isothermal emission model of gas in collisional ionization equilibrium ($kT = 3.0 \pm 0.5$ keV, $\Delta\chi^2 = 129$, d.o.f. = 3 relative to the continuum-only fits). This is able to reproduce the Ne X and O VIII lines and the residual Fe XVII emission, but the O VII emission is underestimated (see red line in Fig. 4). To model the absorption features, we applied a photoionized absorber to the continuum components (blackbody and power-law) of NGC 5408 X-1, leaving the continuum components of X-2 unabsorbed. Most features can be reproduced with a relativistically outflowing ($v = (0.22 \pm 0.01)c$, $\Delta\chi^2 = 35$, d.o.f. = 3) photoionized gas model. A better description of the spectrum is obtained adding a cooler CIE ($kT = 0.10 \pm 0.05$ keV, $\Delta\chi^2 = 15$, d.o.f. = 2) to fit the O VII lines and a slower ($v = (0.10 \pm 0.01)c$, $\Delta\chi^2 = 20$, d.o.f. = 3) photoionized absorber (again only applied to the X-1 continuum components, see blue line in Fig. 4). The absorption lines have a width of $500 \pm 300$ km s$^{-1}$, while the emission lines are broader with $\sigma_{\rm v} = 2,000 \pm 500$ km s$^{-1}$, which are similar to the resolved RGS lines in NGC 1313 X-1. Solar abundances were adopted for all emission and absorption components. The highly significant, $v = 0.22c$, absorber in NGC 5408 X-1 has an ionization parameter $\xi = 50 \pm 30$ erg cm s$^{-1}$ and column density $N_{\rm H} = (3.0 \pm 0.4) \times 10^{20}$ cm$^{-2}$, which are lower than in NGC 1313 X-1, while their outflow velocities ($v \approx 0.2c$) are comparable.

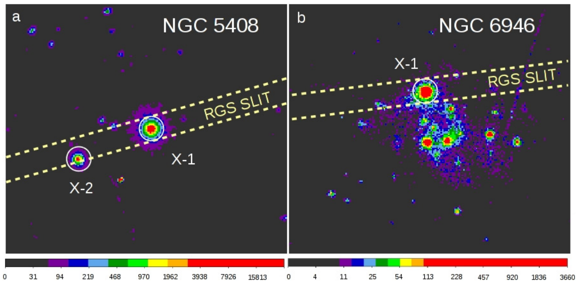

20. Strüder, L. *et al.* The European Photon Imaging Camera on XMM-Newton: the pn-CCD camera. *Astron. Astrophys.* **365,** L18–L26 (2001).
21. Turner, M. J. L. *et al.* The European Photon Imaging Camera on XMM-Newton: the MOS cameras. *Astron. Astrophys.* **365,** L27–L35 (2001).
22. den Herder, J. W. *et al.* The Reflection Grating Spectrometer on board XMM-Newton. *Astron. Astrophys.* **365,** L7–L17 (2001).
23. Roberts, T. P. *et al.* Chandra monitoring observations of the ultraluminous X-ray source NGC 5204 X-1. *Mon. Not. R. Astron. Soc.* **371,** 1877–1890 (2006).
24. Kalberla, P. M. W. *et al.* The Leiden/Argentine/Bonn (LAB) Survey of Galactic HI. Final data release of the combined LDS and IAR surveys with improved stray-radiation corrections. *Astron. Astrophys.* **440,** 775–782 (2005).
25. Lodders, K. S. & Palme, H. Solar system elemental abundances in 2009. *Meteorit. Planet. Sci. Suppl.* **72,** 5154 (2009).
26. Nomoto, K. *et al.* Nucleosynthesis yields of core-collapse supernovae and hypernovae, and galactic chemical evolution. *Nucl. Phys.* **777,** 424–458 (2006).
27. Swartz, D. A. *et al.* The ultraluminous X-ray source population from the Chandra archive of galaxies. *Astrophys. J. Suppl. Ser.* **154,** 519–539 (2004).
28. Pakull, M. W. & Mirioni, L. Optical counterparts of ultraluminous X-ray sources. Preprint at http://arxiv.org/abs/astro-ph/0202488 (2002).
29. Goad, M. R., Roberts, T. P., Reeves, J. N. & Uttley, P. A deep XMM-Newton observation of the ultraluminous X-ray source Holmberg II X-1: the case against a 1000-$M_\odot$ black hole. *Mon. Not. R. Astron. Soc.* **365,** 191–198 (2006).
30. Pinto, C., Kaastra, J. S., Costantini, E. & de Vries, C. Interstellar medium composition through X-ray spectroscopy of low-mass X-ray binaries. *Astron. Astrophys.* **551,** A25–A35 (2013).
31. Kaastra, J. S. *et al.* Multiwavelength campaign on Mrk 509. II. Analysis of high-quality Reflection Grating Spectrometer spectra. *Astron. Astrophys.* **534,** A37–A52 (2011).
32. King, A. L. *et al.* Regulation of black hole winds and jets across the mass scale. *Astrophys. J.* **762,** 103–120 (2013).



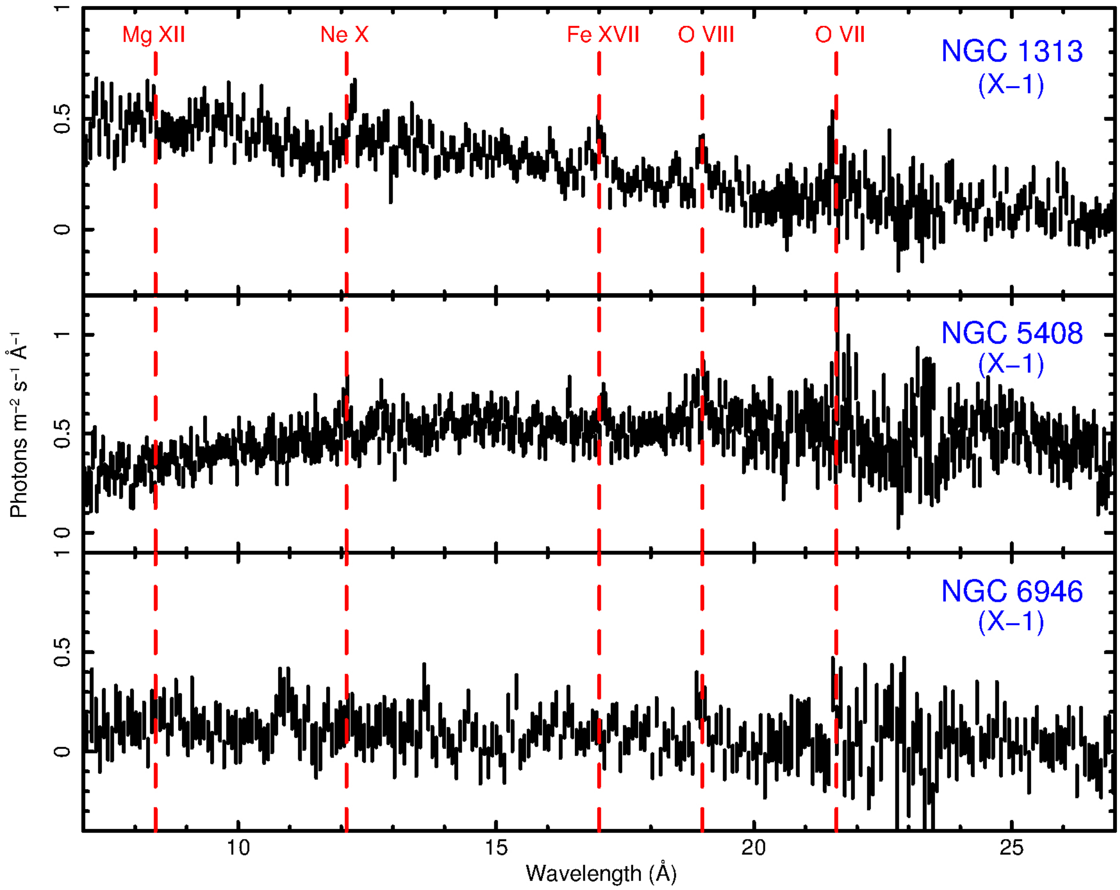

**Extended Data Figure 1 | EPIC MOS+PN stacked image of NGC 1313.** The circular source extraction regions (large white circles) have a diameter of 1 arcmin. The small region to the south of X-1 (small white circle) is a star-forming region near the galactic centre, orders of magnitude fainter across the 0.3–10 keV bandpass than X-1. The strip enclosed within dashed yellow lines is the RGS extraction region. Counts per pixel are colour coded (key at bottom).



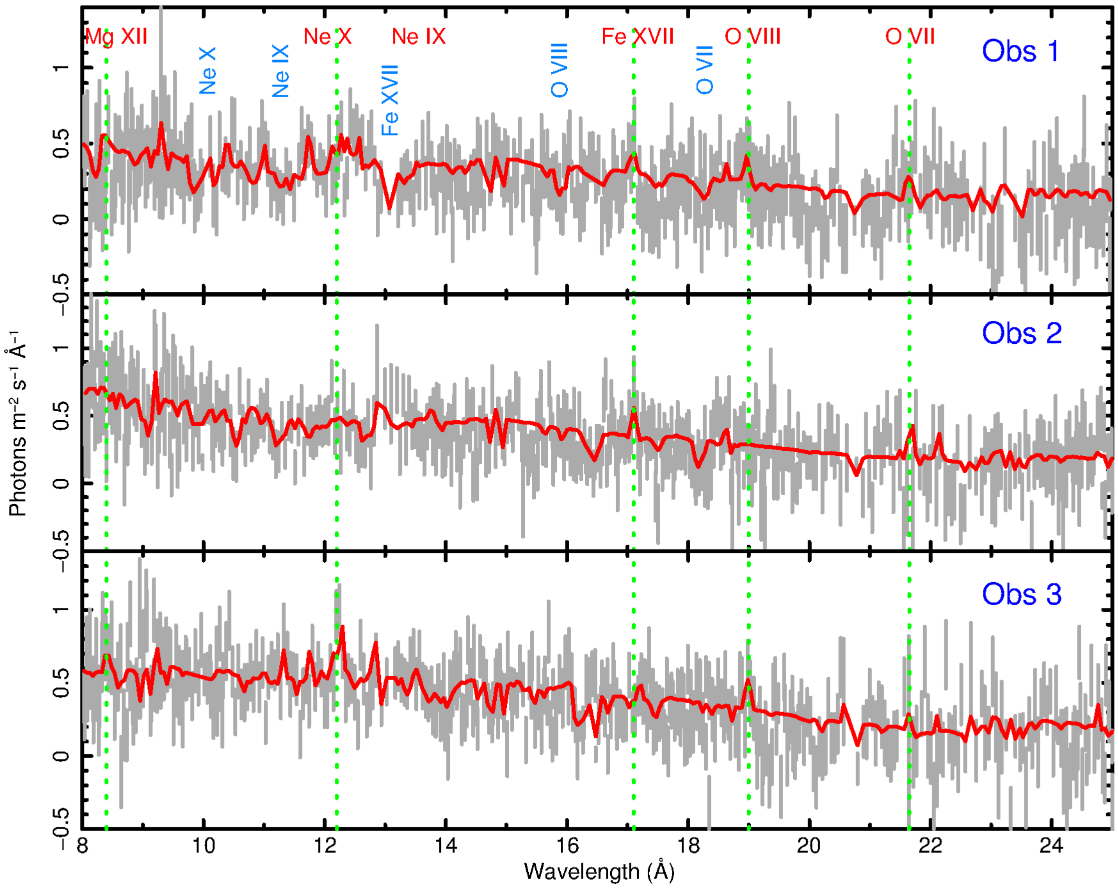

**Extended Data Figure 2 | ACIS image of NGC 1313 X-1 and the nearby star-forming region, SFR.** The ultraluminous X-ray source is the brightest object. The small circles have 6 arcsec radii, that is, 0.1 arcmin; the larger circle has 0.5 arcmin radius. Counts per pixel are colour coded (key at bottom).



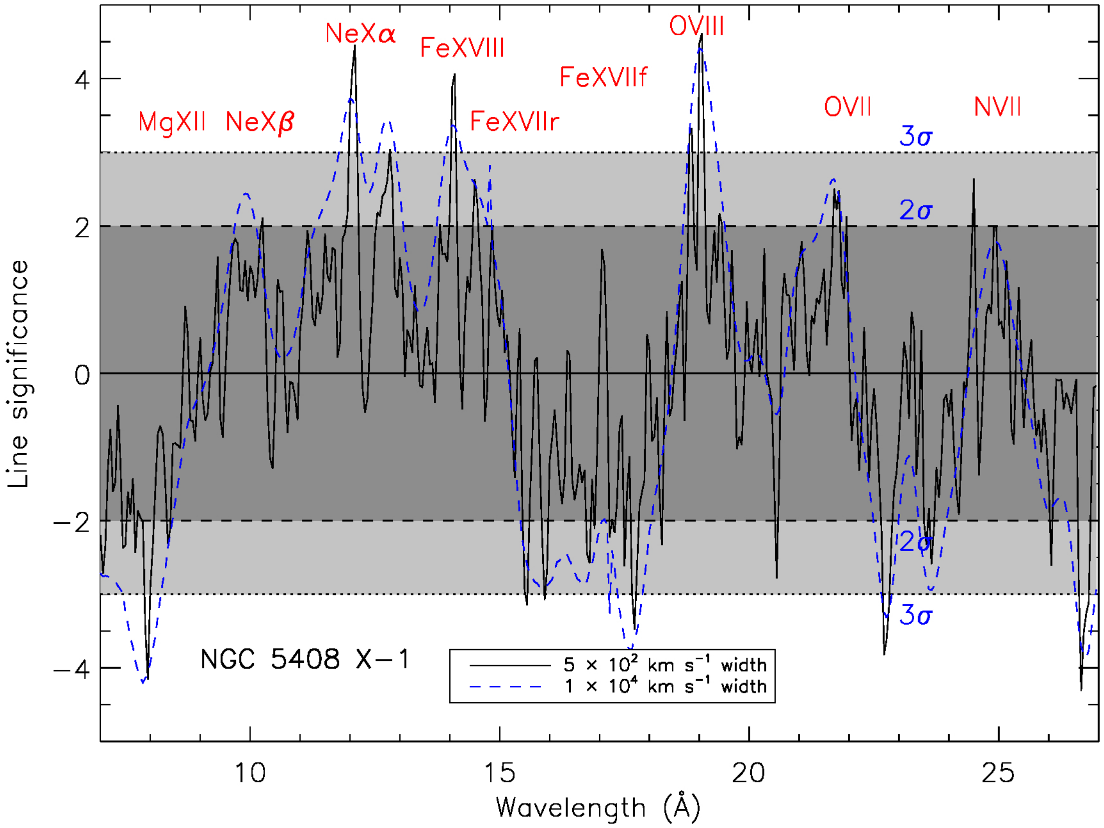

**Extended Data Figure 3 | EPIC MOS+PN stacked images of NGC 5408 and NGC 6946. a**, NGC 5408; **b**, NGC 6946. The ultraluminous X-ray sources are the brightest objects in both images. Additional, nearby X-ray bright sources—mostly high-mass X-ray binaries and background active galactic nuclei—can be seen. The white circular source extraction regions have a diameter of 1 arcmin. Counts per pixel are colour coded (key at bottom).



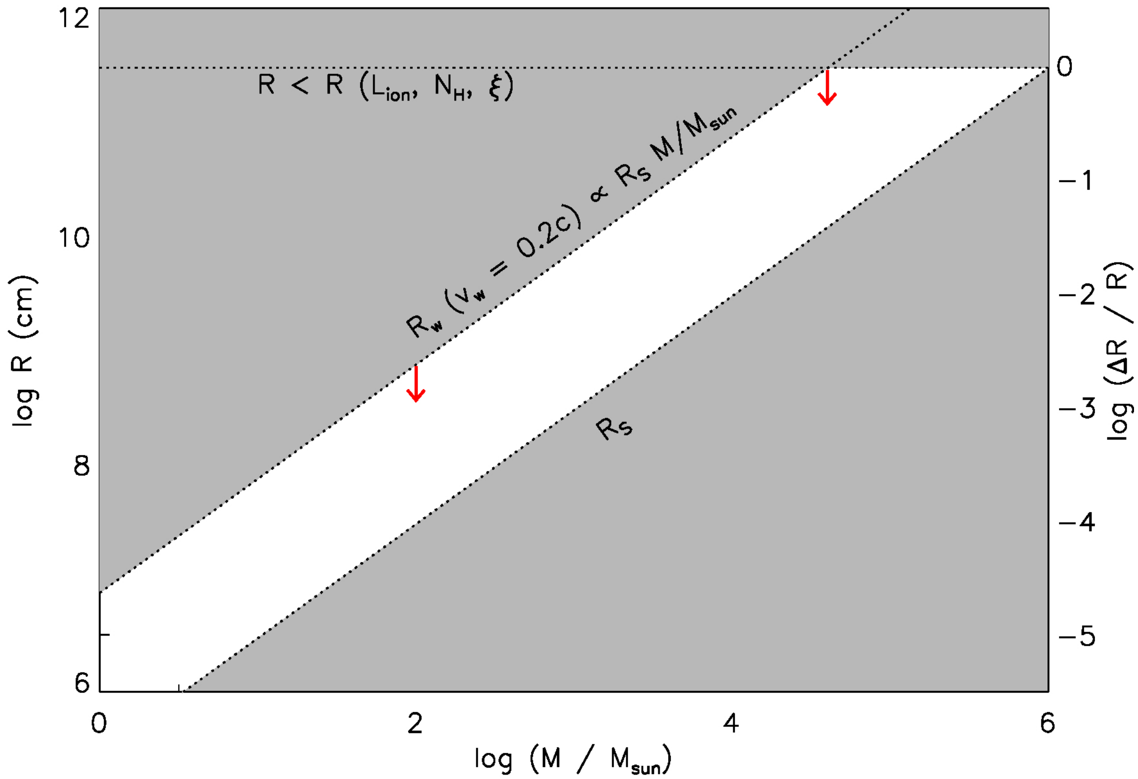

**Extended Data Figure 4 | XMM-Newton/RGS stacked spectra of the brightest ULXs (X-1) in NGC 1313, NGC 5408 and NGC 6946.** Spectra are in flux units. The rest-frame wavelengths of relevant transitions are given as vertical red dashed lines, labelled with the transition. The spectra have been re-binned for display purposes. Error bars, $\pm 1\sigma$.




**Extended Data Figure 5 | XMM-Newton RGS spectra and best-fitting model to each observation of NGC 1313 X-1.** Spectra are in flux units. The rest-frame wavelengths of the most relevant emission lines (green vertical dashed lines) are shown, labelled with the transition in red, and the rest-frame wavelengths of the blueshifted absorption lines are shown by the position of the transition in blue. Obs 1 shows both absorption and emission lines. Obs 2 is dominated by absorption, while obs 3 shows mostly emission features. Error bars, $\pm 1\sigma$.




**Extended Data Figure 6 | Significance of the features in the NGC 5408 X-1 RGS stacked spectrum.** Negative values refer to absorption lines (see also Fig. 3 for NGC 1313 X-1). The solid and dashed curves show the line significance obtained with 500 km s$^{-1}$ and 10,000 km s$^{-1}$ widths, respectively. The dark and light grey regions enclose points within 2$\sigma$ and 3$\sigma$ confidence levels, respectively. Relevant transitions are labelled in red.




**Extended Data Figure 7 | Constraints on the location of the extreme absorber, XABS 3, and mass of the compact object.** The white area shows the acceptable values between the Schwarzschild radius ($R_S$, bottom oblique line), the relation $R = f(L_{ion}, N_H, \xi)$, which is given by $\xi = L/n_H R^2$ (dotted horizontal line), and the radius assuming the escape velocity to be equal to the wind speed ($v_w = 0.2c$, top oblique line, where $M$ and $M_{sun}$ are the compact object and solar masses, respectively). The red arrows show the maximum radius and the upper limit for a compact object with a $100M_{sun}$ mass.

**Extended Data Table 1 | Summary of the XMM-Newton observations**

| Source | Observation ID | $t_{TOT}$ (ks) | $L_{0.3\text{-}10\ keV}$ (erg $s^{-1}$) |
|---|---|---|---|
| NGC 1313 X-1 | 0405090101, 0693850501, 0693851201 | 345.6 | $1.04 \times 10^{40}$ |
| NGC 5408 X-1 | 0302900101, 0500750101, 0653380201, 0653380301, 0653380401, 0653380501 | 644.9 | $2.01 \times 10^{40}$ |
| NGC 6946 X-1 | 0691570101 | 110.0 | $0.97 \times 10^{40}$ |

$t_{TOT}$, exposure times after data reduction; $L_{0.3\text{–}10keV}$, average de-absorbed luminosities.

**Extended Data Table 2 | XMM-Newton EPIC-RGS spectral modelling**

**Model 1 with narrow lines**

| Parameter | CIE | XABS 1 | XABS 2 |
|---|---|---|---|
| $N_{CIE}$ or $N_{H,XABS}$ | $3.1 \pm 0.4$ | $1.5 \pm 0.3 \times 10^{-2}$ | $5 \pm 1 \times 10^{-3}$ |
| $T_{CIE}$ or $\xi_{XABS}$ | $0.85 \pm 0.03$ | $2.20 \pm 0.04$ | 2.20 coupled |
| $v_{\sigma}$ | 10 (< 500) | 10 coupled | 10 coupled |
| $v_{outflow}$ | ≡0 | 0 (> $-3 \times 10^3$) | $-6.5 \pm 0.2 \times 10^4$ |
| O/Fe | $1.0 \pm 0.2$ | 1.0 coupled | 1.0 coupled |
| Ne/Fe | $1.8 \pm 0.4$ | 1.8 coupled | 1.8 coupled |
| Mg/Fe | $2.1 \pm 0.5$ | 2.1 coupled | 2.1 coupled |
| $\Delta\chi^2$, $\Delta C_{stat}$, *d.o.f.* | 87, 114, 6 | 130, 65, 3 | 48, 20, 3 |

**Model 2 with broad lines**

| Parameter | CIE | XABS 1 | XABS 2 | XABS 3 |
|---|---|---|---|---|
| $N_{CIE}$ or $N_{H,XABS}$ | $3.0 \pm 0.4$ | $3.8 \pm 1.3 \times 10^{-3}$ | $3.1 \pm 1.0 \times 10^{-3}$ | $1.1 \pm 0.2$ |
| $T_{CIE}$ or $\xi_{XABS}$ | $0.80 \pm 0.03$ | $2.29 \pm 0.09$ | 2.29 coupled | $4.55 \pm 0.22$ |
| $v_{\sigma}$ | $1250 \pm 600$ | 10 (< 20) | 10 (< 20) | $3.0 \pm 1.5 \times 10^4$ |
| $v_{outflow}$ | ≡0 | 0 (> $-3 \times 10^3$) | $-3.9 \pm 0.2 \times 10^4$ | $-7.5 \pm 1.5 \times 10^4$ |
| O/Fe | $1.38 \pm 0.16$ | 1.38 coupled | 1.38 coupled | 1.38 coupled |
| Ne/Fe | $3.88 \pm 0.91$ | 3.87 coupled | 3.87 coupled | 3.87 coupled |
| Mg/Fe | $3.33 \pm 0.50$ | 3.33 coupled | 3.33 coupled | 3.33 coupled |
| $\Delta\chi^2$, $\Delta C_{stat}$, *d.o.f.* | 86, 107, 6 | 30, 15, 4 | 12, 10, 4 | 150, 41, 4 |

The table is divided in two blocks to detail the results obtained with Model 1 (narrow absorption and emission lines) and Model 2 (Model 1 with an additional velocity-broadened absorption component). The CIE normalizations $N_{CIE}$ (0.3–10 keV luminosity) and temperatures $T_{CIE}$ (in $kT$, keV, units) are in units of $10^{38}$ erg s$^{-1}$ and keV, respectively; the XABS column densities ($N_{H,XABS}$) and ionization parameters ($\xi_{XABS}$) are reported in units of $10^{24}$ cm$^{-2}$ and log[$\xi$(erg cm s$^{-1}$)], respectively; both velocity broadening $v_{\sigma}$ and outflow velocity are in km s$^{-1}$. The abundances are relative to iron, whose abundance is fixed to be solar[25]. Errors are $\pm 1\sigma$.

# LETTER

doi:10.1038/nature17628

# Polar metals by geometric design

T. H. Kim[1], D. Puggioni[2], Y. Yuan[3], L. Xie[4,5], H. Zhou[6], N. Campbell[7], P. J. Ryan[6], Y. Choi[6], J.-W. Kim[6], J. R. Patzner[1], S. Ryu[1], J. P. Podkaminer[1], J. Irwin[7], Y. Ma[1], C. J. Fennie[8], M. S. Rzchowski[7], X. Q. Pan[4], V. Gopalan[3], J. M. Rondinelli[2] & C. B. Eom[1]

**Gauss's law dictates that the net electric field inside a conductor in electrostatic equilibrium is zero by effective charge screening; free carriers within a metal eliminate internal dipoles that may arise owing to asymmetric charge distributions[1]. Quantum physics supports this view[2], demonstrating that delocalized electrons make a static macroscopic polarization, an ill-defined quantity in metals[3]—it is exceedingly unusual to find a polar metal that exhibits long-range ordered dipoles owing to cooperative atomic displacements aligned from dipolar interactions as in insulating phases[4]. Here we describe the quantum mechanical design and experimental realization of room-temperature polar metals in thin-film $ANiO_3$ perovskite nickelates using a strategy based on atomic-scale control of inversion-preserving (centric) displacements[5]. We predict with *ab initio* calculations that cooperative polar $A$ cation displacements are geometrically stabilized with a non-equilibrium amplitude and tilt pattern of the corner-connected $NiO_6$ octahedra—the structural signatures of perovskites—owing to geometric constraints imposed by the underlying substrate. Heteroepitaxial thin-films grown on $LaAlO_3$ (111) substrates fulfil the design principles. We achieve both a conducting polar monoclinic oxide that is inaccessible in compositionally identical films grown on (001) substrates, and observe a hidden, previously unreported[6–10], non-equilibrium structure in thin-film geometries. We expect that the geometric stabilization approach will provide novel avenues for realizing new multifunctional materials with unusual coexisting properties.**

Polar metals are characterized by the absence of inversion symmetry and are intrinsically conducting owing to partial band occupation. This operational definition excludes degenerately doped insulating ferroelectrics. It further exacerbates the complexity of designing crystalline phases that exhibit both a polar crystal structure and Drude-type conductivity. Nonetheless, Anderson proposed more than a half century ago that polar metals could exist[11], and serendipitous discoveries of such materials have occurred[12,13]. A recent example is an osmate requiring high-pressure synthesis[12]. Although polar metals remain scarce, predictive guidelines have been proposed[14]. However, materials with benign chemistries and device-relevant two-dimensional geometries remain to be realized.

A key to designing polar metals is that the electronic structure at the Fermi level is mainly derived from orbital states that are decoupled

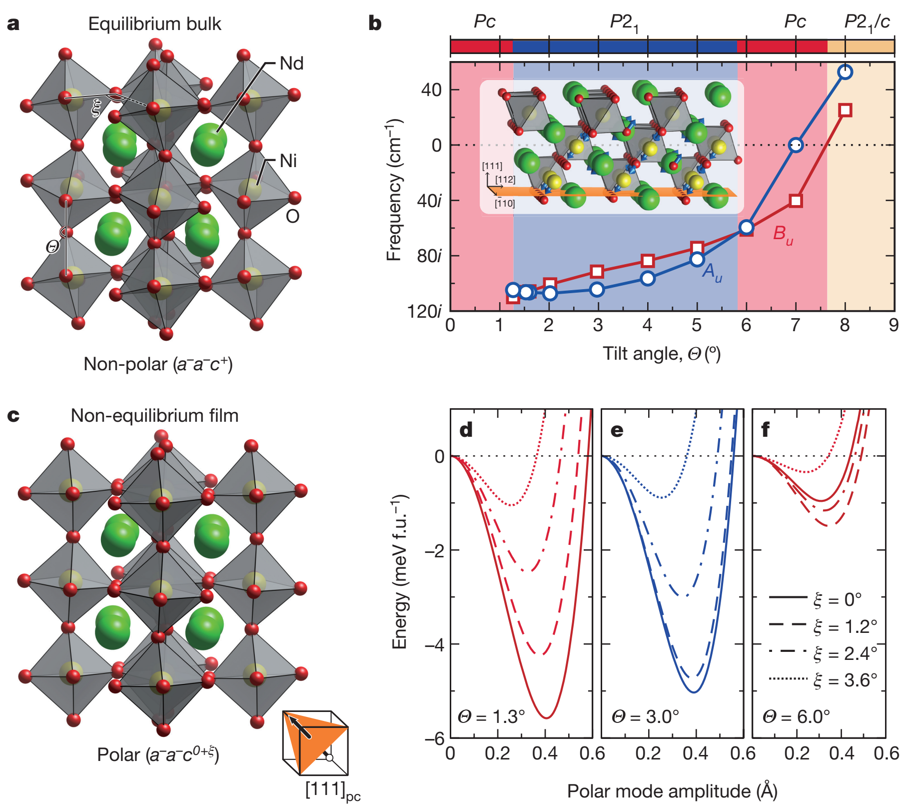



**Figure 1 | Geometric stabilization of polar $NdNiO_3$ via octahedral tilt engineering. a**, The equilibrium room temperature bulk structure of metallic $NdNiO_3$ with the $a^-a^-c^+$ tilt pattern formed by corner-connected $NiO_6$ with tilt ($\Theta$) and rotation ($\xi$) angles. **b**, The calculated zone-centre phonon modes with $A_u$ (blue line) and $B_u$ (red line) symmetry for a centrosymmetric $NdNiO_3$ on $LaAlO_3$ (111) with varying degree of the $a^-a^-c^0$ tilt angle $\Theta$ with the in-phase rotation angle $\xi = 0°$. Imaginary frequencies indicate dynamical lattice instabilities, which harden as the tilt angle $\Theta$ increases towards its equilibrium value. The inset depicts the atomic displacement directions for the $B_u$ mode. The top panel indicates the theoretically predicted stable phases for small tilt angles (<7.6°), above which a non-polar metallic phase is stable. **c**, The non-equilibrium geometrically stabilized polar metal $NdNiO_3$ structure on $LaAlO_3$ (111), depicted in a pseudocubic (pc) representation, for $\Theta \approx 6°$; the reduction in the octahedral tilt angles and rotation pattern is clearly discernible. **d**–**f**, Energetic gain for varying octahedral tilt ($\Theta$) and rotation ($\xi$) angles with the unstable $A_u$ or $B_u$ modes. While the energy gain is maximized for $\xi = 0°$ and $\Theta < 6°$ (**d**, **e**), a small rotation angle $\xi \leq 1.2°$ is needed for $\Theta \geq 6°$ (**f**). Colours and lines follow the scheme in **b**. f.u., formula unit.

[1]Department of Materials Science and Engineering, University of Wisconsin-Madison, Madison, Wisconsin 53706, USA. [2]Department of Materials Science and Engineering, Northwestern University, Evanston, Illinois 60208, USA. [3]Department of Materials Science and Engineering and Materials Research Institute, Pennsylvania State University, University Park, Pennsylvania 16802, USA. [4]Department of Chemical Engineering and Materials Science and Department of Physics and Astronomy, University of California-Irvine, Irvine, California 92697, USA. [5]National Laboratory of Solid State Microstructures and College of Engineering and Applied Sciences, Nanjing University, Nanjing, Jiangsu 210093, China. [6]Advanced Photon Source, Argonne National Laboratory, Argonne, Illinois 60439, USA. [7]Department of Physics, University of Wisconsin-Madison, Madison, Wisconsin 53706, USA. [8]School of Applied and Engineering Physics, Cornell University, Ithaca, New York 14853, USA.

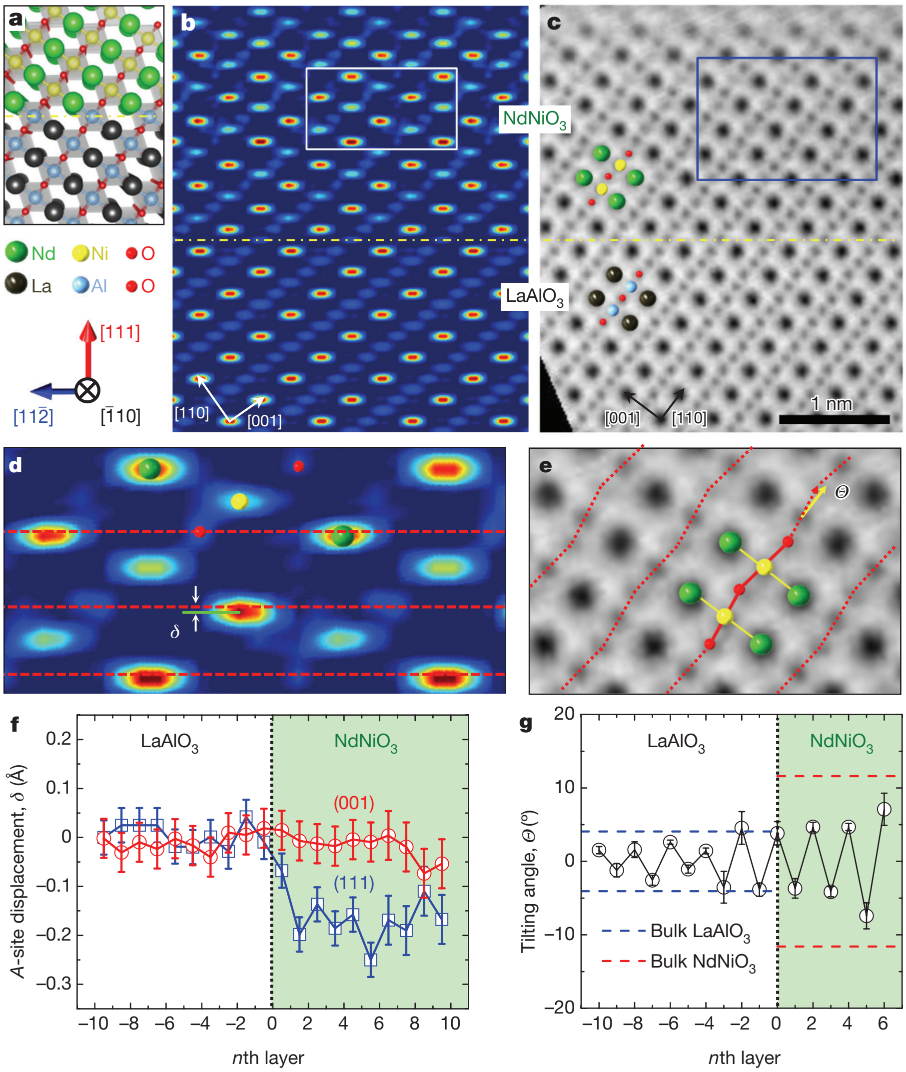



**Figure 2 | Non-centrosymmetric $NdNiO_3$ thin films on $LaAlO_3$ (111) substrates. a**, Schematic illustration of the atomic-scale thin-film heterostructure, **b**, **c**, Two-dimensional electron density maps sliced through the pseudocubic (1$\bar{1}$0) plane reconstructed through synchrotron CTR measurements and subsequent COBRA analyses (**b**) and STEM-ABF images captured along the pseudocubic [$\bar{1}$10] zone axis in cross-sectional view (**c**). In **a**–**c**, the interface is marked by the yellow dash–dotted lines. **d**, **e**, Magnified images of electron density maps (**d**) and ABF images (**e**) for the regions indicated by open rectangles in **b** and **c**, respectively. In **d**, red broken lines represent the positions of oxygen atoms (marked with red colours), which are taken as references to measure relative off-centre displacements ($\delta$) of Nd atoms (marked with green colours). In **e**, red broken lines are used as guidelines to show tilting of the $NiO_6$ octahedra in the $NdNiO_3$ layer with an angle of $\Theta$, obtained by calculating the angle formed between a line (O–O, red dotted arrow) connecting two nearest oxygen atoms and another line (*B*–*B*, yellow solid arrow) connecting two nearest *B*-site atoms. **f**, **g**, Layer-dependent evolution of the *A*-site relative polar displacements (**f**) and $BO_6$ octahedra tilt angles (**g**) across the interface in $NdNiO_3/LaAlO_3$ (111) thin films. The 0th layer represents the $NdNiO_3/LaAlO_3$ interface. In the two-dimensional electron density map of **b**, the *A*-site acentric displacements shown in **f** are measured with respect to oxygen atoms as displayed in **d**. Error bars are statistical, based on measurements of the *A*-site displacement (**f**) or tilting angle (**g**). Details of the statistical analyses used are described in Methods. In **g**, the blue and red broken lines represent the tilting angles of bulk $LaAlO_3$ (~4.06°) and $NdNiO_3$ (~11.6°), respectively.

from the ions undergoing polar atomic displacements. We form this weak electron-lattice coupling principle[11,14] into materials selection criteria as follows. First, select a chemistry to ensure finite band occupation (metallicity). Second, choose a crystal structure compatible with an electronic-structure-insensitive mechanism for polar ionic displacements. We simultaneously satisfy both constraints by selecting ternary $ABO_3$ perovskite oxides. The multiple *A* and *B* cations provide two sublattices, which permit one to contribute partially occupied delocalized states for conduction while the other is able to undergo polar displacements in response to changes in the $BO_6$ metal–oxygen octahedral framework (Fig. 1a).

We test the feasibility of this approach in the rare-earth (*R*) nickelates $RNiO_3$, which undergo thermally driven metal-to-insulator transitions dependent on the crystallographic tolerance factor, *t*, which is an effective measure of the *A* and *B* cation size mismatch. The electronic phase diagram[15] reveals that $NdNiO_3$, $PrNiO_3$ or $LaNiO_3$ with $0.91 \leq t \leq 0.94$ are suitable candidates as they are electrically conductive at room temperature[16]. At equilibrium, all bulk compounds exhibit centrosymmetric structures with either an orthorhombic structure (*Pnma* and $a^-a^-c^+$ tilt pattern sketched in Fig. 1a) or a rhombohedral structure ($R\bar{3}c$, $a^-a^-a^-$ tilt as for $LaNiO_3$).

We now address the second criterion: a driving force for inversion-lifting displacements from a set of atoms decoupled from the low-energy electronic structure. Recently, it has been theoretically shown[17] that orthorhombic perovskites would be susceptible to polar *A*-site (*B* and O) displacements if the anti-polar *R* cation displacements induced by the $a^-a^-c^+$ rotations (Fig. 1a) were suppressed[18,19], for example, either by reducing one or more of the $BO_6$ rotations or by stabilizing the $a^-a^-a^-$ rotation pattern. Importantly, the electrons derived from the $R^{3+}$ cations are independent of the conduction in the nickelates. Thus, if the $RNiO_3$ compounds could be realized in a non-equilibrium structure, that is, with a non-equilibrium tilt pattern (a change in the rotation and tilt amplitude and/or sense), then polar *R* displacements could be achieved.

As polar displacements are more favourable in smaller tolerance factor perovskites, we first select $NdNiO_3$ and then examine $LaNiO_3$ (with a larger *t*) later. We first use first-principles calculations to evaluate the structure (tilt) stability of $NdNiO_3$ under the epitaxial boundary conditions imposed by the isoelectronic substrate $LaAlO_3$ with pseudocubic (001) and (111) orientations. The different orientations lead to a change in the number of Ni–O–Al bond contacts across the heterointerface: the (001) and (111) orientations permit a single and three Al–O–Ni bond connections along the [001] and [111] directions, respectively, (Extended Data Fig. 1). The (111) heterointerface enables greater control over the $NiO_6$ tilt pattern and electronic properties[20] by coupling of the interfacial octahedra through substrate proximity effects[21]. We find that the bulk orthorhombic $a^-a^-c^+$ tilt is energetically more favourable than any other tilt pattern explored for both substrate orientations (Extended Data Table 1). Focusing on the (111) case, we find the low-energy structure to be a centrosymmetric phase: $P2_1/c$ symmetry with the $a^-a^-c^+$ tilt pattern with tilt ($\Theta$) and rotation ($\xi$) angles of 11.6° and 7.0°, respectively. A non-equilibrium $NiO_6$ tilt pattern is only accessible through the geometric constraint imposed by the underlying substrate orientation. Thus, a non-centrosymmetric metallic state may

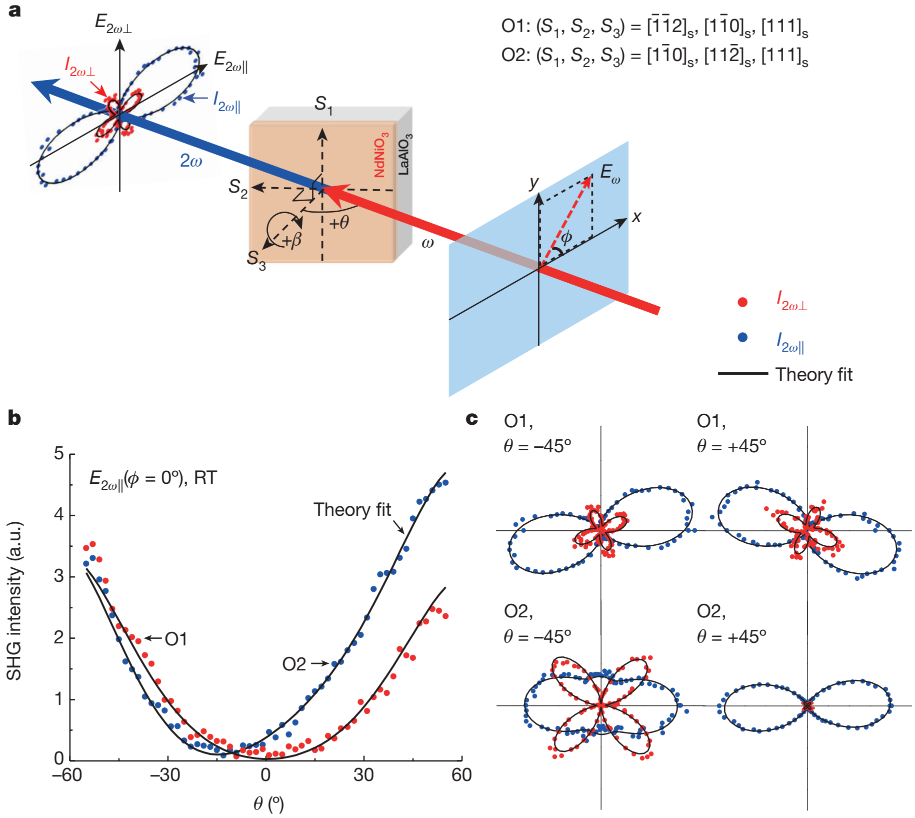



**Figure 3 | SHG polarimetry of $NdNiO_3$ (111) thin films. a**, Schematic diagram of SHG measurements with respect to two different sample orientations (O1 and O2), which are described by the crystallographic directions of the $LaAlO_3$ substrate (subscript s), as indicated. **b**, **c**, The incident ($\theta$) (**b**) and polarization ($\phi$) (**c**) angle dependence of the room-temperature (RT) SHG response. In **b**, two SHG components are plotted, $I_{2\omega||}$ (blue) and $I_{2\omega\perp}$ (red), with analysers parallel and perpendicular to the incidence plane, respectively. The ten sets of SHG data in **b** and **c** are simultaneously fitted with the same parameters to the point group symmetry *m* with three equivalent domain variants. a.u., arbitrary units.

be achieved from substrate constraints that reduce the amplitude of the tilt angles in the $P2_1/c$ phases.

To assess the feasibility of a polar $NdNiO_3$ state arising from such interactions, we compute the zone-centre phonon modes in the more-stable structure as a function of tilt angle with the in-phase rotation angle $\xi = 0°$ (Fig. 1b). We find two polar modes with $B_u$ and $A_u$ symmetry, resulting in polar *Pc* (Fig. 1c) and chiral-polar $P2_1$ symmetries, respectively, owing to ionic displacements on the $(1\bar{1}0)_{pc}$ (*Pc*) and $(001)_{pc}$ ($P2_1$) planes. Each mode frequency hardens as the $NiO_6$ tilt angle $\Theta$ increases, becoming energetically unfavourable for $\Theta \gtrsim 7.6°$ ($B_u$) and $\Theta \gtrsim 7.0°$ ($A_u$), above which the centric $P2_1/c$ is stable (Fig. 1b). We find that the polar structures are most stable for tilt angles $\Theta \leq 1.5°$ and $5.8° \leq \Theta \leq 7.6°$ (Fig. 1b), and the stability of these polar phases vanishes upon recovering the equilibrium tilt pattern (Fig. 1d–f). Similar structure stability results are obtained assuming an equilibrium tilt pattern $a^-a^-c^-$ ($C2/c$), which should be exhibited by the $LaNiO_3$ film on the $LaAlO_3$ (111) substrate (Extended Data Fig. 2).

Taken together, our first-principles calculations predict that polar displacements will follow if a non-equilibrium $NiO_6$ tilt pattern and angles can be stabilized in thin-film $NdNiO_3$. To activate the required tilt pattern, we experimentally synthesized high-quality $NdNiO_3$ films using pulsed laser deposition with *in situ* reflection high-energy electron diffraction (RHEED) on the $LaAlO_3$ (111) substrate (Extended Data Fig. 3), which provides a stronger geometric constraint compared with the (001) substrate owing to increased bond connectivity. To determine the polar nature of epitaxial $NdNiO_3$ thin films, various experimental methods are exploited: we quantify the Nd and Ni displacements with synchrotron crystal truncation rod (CTR) measurements and coherent Bragg rod analysis (COBRA), $NiO_6$ tilt-angle suppression with high-resolution scanning transmission electron microscopy (STEM), and macroscopic polar point-group symmetry with second harmonic generation (SHG) polarimetry.

Figure 2 shows the polar Nd displacements ($\delta$) in the (111) film. The *A*-site (Nd and La), *B*-site (Ni and Al), and oxygen atoms appear in the COBRA two-dimensional electron density map sliced through the ⟨110⟩ planes (Fig. 2b). Across the heterointerface, acentric Nd displacements are pronounced with respect to oxygen atoms along $[111]_{pc}$ (Fig. 2d, f), which is in agreement with our theoretical calculations for the *Pc* structures. The cooperative polar Nd and Ni displacements were further verified using high-resolution synchrotron X-ray diffraction (Extended Data Fig. 3e) and near-edge X-ray absorption fine structure (NEAXFS) (Extended Data Fig. 4). In contrast, no resolvable polar displacements were observed in the reconstructed electron density map of an $NdNiO_3$ (001) film (Fig. 2f and Extended Data Fig. 5), indicating that a key part is played by the interfacial geometry.

Next, we demonstrate that the polar displacements arise from the local suppression of a $NiO_6$ tilt angle. The tilting angle ($\Theta$) in a STEM annular bright field (ABF) image is measured by a line connecting two nearest oxygen atoms and intersecting a line connecting two nearest *B*-site cations (Fig. 2c, e). It is evident that the $NiO_6$ tilting angle varying between 4° and 8° in the (111) film is reduced from the bulk value (11.6°) (Fig. 2g), as proposed by our theoretically predicted geometric stabilization strategy. A detailed analysis of the octahedral tilting leads us to conclude that the bulk equilibrium $a^-a^-c^+$ tilt pattern is suppressed, leading to the low-energy $a^-a^-c^{0+\xi}$ tilt predicted from first principles with subtle local variations of the in-phase rotation amplitude $0° \leq \xi \leq 1.2°$ (Fig. 1f, Extended Data Fig. 6a–c). In contrast, $NdNiO_3$ (001) films exhibited the orthorhombic $a^-a^-c^+$ tilt pattern (Extended Data Fig. 6d–h) without suppression of the $NiO_6$ tilt angle (Extended Data Fig. 5h), and hence no predicted polar state. The observed tilt patterns in the STEM analyses of both $NdNiO_3$ (111) and (001) thin films are confirmed by synchrotron X-ray diffraction measurements of characteristic half-order Bragg peaks due to octahedral tilting.

To quantify the point-group symmetry of the $NdNiO_3$ (111) films, we performed optical SHG polarimetry analyses. Temperature-dependent SHG responses were also recorded as a function of the incident ($\theta$) and azimuthal ($\phi$) angles of the incident linear polarized beam ($E_\omega$) with the film orientations described in Fig. 3a. Ten sets of polarimetry data in Fig. 3b, c were fitted simultaneously with the same parameter set to

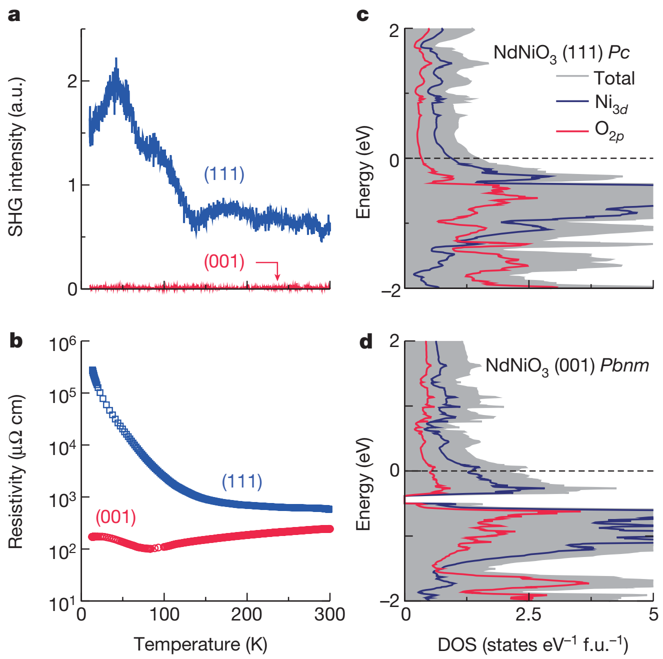



**Figure 4 | Coexistence of polar displacements and metallic conductivity in $NdNO_3$ (111) thin films. a, b,** The temperature-dependent SHG responses (**a**) and resistivity (**b**) of $NdNiO_3$ (111) (2.2 nm) and (001) (3.8 nm) thin films grown on $LaAlO_3$ substrates. The SHG signals and electrical transport properties of $NdNiO_3$ (111) (blue) and (001) (red) thin films were measured during a cooling process. **c, d,** The total (grey area), Ni 3*d* (blue line), and O 2*p* (red line) resolved densities of states of $NdNiO_3$ for the *Pc* ((111) geometry; $\Theta = 6°$ and $\xi = 1.2°$) (**c**) and *Pbnm* ((001) geometry) (**d**) symmetries, respectively. The Fermi level is at 0 eV (dashed line).

various models constructed using non-centrosymmetric point groups of 3*m*, 2, or *m* (see Methods). We find that the best fit to the data is given by the polar monoclinic symmetry of *m* (Fig. 3b, c), which is in good agreement with the predicted *Pc* phase belonging to the polar point group of *m*.

Figure 4 shows the coexistence of polar displacements and metallic conductivity in $NdNiO_3$ (111) films. At all temperatures, polar SHG responses were measured in these films with varying thickness (Extended Data Fig. 7), whereas no SHG signal was detected in $NdNiO_3$ (001) thin films (Fig. 4a). The absence of any SHG signal above the noise floor from the (001)-oriented sample, with the same film and substrate composition as the (111) film and differing only in its crystal orientation, indicates that naturally symmetry-breaking surfaces and interfaces are not responsible for the SHG signal from the (111) film. Furthermore, we observe a quadratic scaling of the SHG intensity as a function of film thickness in the (111) film, which again confirms that the SHG signal arises from the bulk of the film (Extended Data Fig. 7d). Finally, the SHG symmetry analysis reveals a monoclinic point group, consistent with the bulk prediction for this film. We also eliminate the possibility of polar discontinuity at the $NdNiO_3/LaAlO_3$ interface as being the origin of SHG, because all relevant cations across the film/substrate interface are isovalent. Interestingly, a concomitant enhancement of the SHG response occurs for the (111) geometry (Fig. 4a), while both (111) and (001) nickelate films exhibit a metal–insulator transition (Fig. 4b), albeit renormalized from the bulk value. The polar monoclinic symmetry also remains at low temperatures (Extended Data Fig. 7). Previous work indicates that the metal–insulator transition should be suppressed for compressively strained $NdNiO_3$ films grown on $LaAlO_3$ (001)[22]; however, our films are much thinner (10 unit cells (3.8 nm) thick) and are expected to have an enhanced effective mass[23]. The semiconducting transport behaviour in (111) $NdNiO_3$ is also attributed to ultrathin film thickness (Extended Data Fig. 7e). Despite the suppressed electrical conductivity, it is very interesting that the films show polar behaviour in the presence of free carriers. As a consequence, the insulating $NdNiO_3$ is a polar antiferromagnetic at low temperatures, which makes it a potential multiferroic.

Although the $NdNiO_3$ (111) films are less conductive than the (001)-oriented films at all temperatures, our Hall measurements confirm the presence of mobile charge carriers, whose concentration is sufficient to produce metallic conductivity at room temperature. The measured carrier concentration of the $NdNiO_3$ (111) films is $\sim 2.0 \times 10^{20}\,cm^{-3}$, which is less than that of the $NdNiO_3$ (001) films ($\sim 1.5 \times 10^{21}\,cm^{-3}$). This difference in the carrier concentration is consistent with the calculated electronic density of states (DOS) shown in Fig. 4c and d, respectively. Both structures are metallic with a finite DOS at the Fermi level derived from the Ni 3*d* states, with the different spectral weight at the Fermi level supporting the carrier concentrations obtained from Hall measurements.

Next, we consider driving $LaNiO_3$, with its larger tolerance factor, into a polar metallic state through octahedral control in a (111)-oriented thin film. We find that even the more conducting $LaNiO_3$ (111) films also exhibit polar structures at all temperatures (Extended Data Fig. 8a–c). These results further indicate that the polar metallic state may rely on geometric stabilization of non-equilibrium structures. Extrinsic effects arising from interface chemical intermixing (Fig. 2b), point and dipole defects, and surface effects[24] (Extended Data Fig. 8d) are also eliminated as possible origins for the polar displacements. Thus, we conclude that octahedral tilt control in polyhedral structured oxides is a useful strategy to realize contraindicated functionalities in non-equilibrium phases.

The genesis for the stability of the polar metallic state in thin-film $RNiO_3$ is the non-equilibrium octahedral tilt pattern obtained in the geometrically constrained heterostructures. We have experimentally demonstrated this theoretical prediction by resolving the tilt patterns, polar displacements and non-centrosymmetric point groups in epitaxial films, and find that the stronger substrate constraint in the (111) orientation is the active mechanism. The approach applied here is an emerging route by which to accelerate the discovery of new multifunctional materials with unusual coexisting properties, such as anisotropic thermoelectric responses and magnetoelectric multiferroics, paving the way for a new generation of devices with the ability to perform simultaneous electrical, magnetic and optical functions. We anticipate our approach for geometric stabilization of a non-equilibrium states to be a fertile platform also for the rational discovery of unconventional ferroic orders[25,26], spin textures[27,28] and topological phases[29,30] in complex oxides, where broken inversion symmetry can be designed into systems without time reversal symmetry and/or strong spin–orbit interactions.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 14 September 2015; accepted 22 February 2016.**
**Published online 20 April 2016.**

1. Jackson, J. *Classical Electrodynamics* 3rd edn (John Wiley & Sons, 1999).
2. Cohen-Tannoudji, C., Dupont-Roc, J. & Grynberg, G. *Photons and Atoms: Introduction to Quantum Electrodynamics* (John Wiley & Sons, 1989).
3. Resta, R. Why are insulators insulating and metals conducting? *J. Phys. Condens. Matter* **14,** R625 (2002).
4. Cohen, R. E. Origin of ferroelectricity in perovskite oxides. *Nature* **358,** 136–138 (1992).
5. Rondinelli, J. M., May, S. J. & Freeland, J. W. Control of octahedral connectivity in perovskite oxide heterostructures: an emerging route to multifunctional materials discovery. *MRS Bull.* **37,** 261–270 (2012).
6. Breckenfeld, E., Chen, Z., Damodaran, A. R. & Martin, L. W. Effects of nonequilibrium growth, nonstoichiometry, and film orientation on the metal-to-insulator transition in $NdNiO_3$ thin films. *ACS Appl. Mater. Interfaces* **6,** 22436–22444 (2014).
7. Middey, S. *et al.* Polarity compensation in ultra-thin films of complex oxides: the case of a perovskite nickelate. *Sci. Rep.* **4,** 6819 (2014).
8. Gibert, M., Zubko, P., Scherwitzl, R., Iniguez, J. & Triscone, J.-M. Exchange bias in $LaNiO_3/LaMnO_3$ superlattices. *Nature Mater.* **11,** 195–198 (2012).
9. Hepting, M. *et al.* Tunable charge and spin order in $PrNiO_3$ thin films and superlattices. *Phys. Rev. Lett.* **113,** 227206 (2014).

10. Bousquet, E. *et al.* Improper ferroelectricity in perovskite oxide artificial superlattices. *Nature* **452,** 732–736 (2008).
11. Anderson, P. W. & Blount, E. I. Symmetry considerations on martensitic transformations: "ferroelectric" metals? *Phys. Rev. Lett.* **14,** 217–219 (1965).
12. Shi, Y. *et al.* A ferroelectric-like structural transition in a metal. *Nature Mater*. **12,** 1024–1027 (2013).
13. Vaughey, J. T. *et al.* Synthesis and structure of a new family of cuprate superconductors: $LnSr_2Cu_2GaO_7$. *Chem. Mater.* **3,** 935–940 (1991).
14. Puggioni, D. & Rondinelli, J. M. Designing a robustly metallic noncentrosymmetric ruthenate oxide with large thermopower anisotropy. *Nature Commun*. **5,** 3432 (2013).
15. Catalan, G. Progress in perovskite nickelate research. *Phase Transit.* **81,** 729–749 (2008).
16. Park, H., Millis, A. J. & Marianetti, C. A. Total energy calculations using DFT+DMFT: computing the pressure phase diagram of the rare earth nickelates. *Phys. Rev. B* **89,** 245133 (2014).
17. Benedek, N. A. & Fennie, C. J. Why are there so few perovskite ferroelectrics? *J. Phys. Chem. C* **117,** 13339–13349 (2013).
18. Bilc, D. I. & Singh, D. J. Frustration of tilts and A-site driven ferroelectricity in $KNbO_3$–$LiNbO_3$ alloys. *Phys. Rev. Lett.* **96,** 147602 (2006).
19. Halilov, S. V., Fornari, M. & Singh, D. J. Lattice instabilities and ferroelectricity in $AScO_3$ perovskite alloys. *Phys. Rev. B* **69,** 174107 (2004).
20. Catalano, S. *et al.* Tailoring the electronic transitions of $NdNiO_3$ films through $(111)_{pc}$ oriented interfaces. *APL Mater.* **3,** 062506 (2015).
21. Rondinelli, J. M. & Spaldin, N. A. Substrate coherency driven octahedral rotations in perovskite oxide films. *Phys. Rev. B* **82,** 113402 (2010).
22. Liu, J. *et al.* Strain-mediated metal-insulator transition in epitaxial ultrathin films of $NdNiO_3$. *Appl. Phys. Lett.* **96,** 233110 (2010).
23. King, P. D. C. *et al.* Atomic-scale control of competing electronic phases in ultrathin $LaNiO_3$. *Nature Nanotechnol*. **9,** 443–447 (2014).
24. Kumah, D. P. *et al.* Tuning the structure of nickelates to achieve two-dimensional electron conduction. *Adv. Mater.* **26,** 1935–1940 (2014).
25. Pitcher, M. J. *et al.* Tilt engineering of spontaneous polarization and magnetization above 300 K in a bulk layered perovskite. *Science* **347,** 420–424 (2015).
26. Oh, Y. S., Luo, X., Huang, F.-T., Wang, Y. & Cheong, S.-W. Experimental demonstration of hybrid improper ferroelectricity and the presence of abundant charged walls in $(Ca,Sr)_3Ti_2O_7$ crystals. *Nature Mater*. **14,** 407–413 (2015).
27. Jonietz, F. *et al.* Spin transfer torques in MnSi at ultralow current densities. *Science* **330,** 1648–1651 (2010).
28. Seki, S., Yu, X. Z., Ishiwata, S. & Tokura, Y. Observation of skyrmions in multeferroic material. *Science* **336,** 198–201 (2012).
29. Rüegg, A. & Fiete, G. A. Topological insulators from complex orbital order in transition-metal oxides heterostructures. *Phys. Rev. B* **84,** 201103 (2011).
30. Okamoto, S. *et al.* Correlation effects in (111) bilayers of perovskite transition-metal oxides. *Phys. Rev. B* **89,** 195121 (2014).


**Acknowledgements** This work was supported by the National Science Foundation (NSF) under Designing Materials to Revolutionize and Engineer our Future grant number DMR-1234096. Transport measurement at the University of Wisconsin–Madison was supported by the US Department of Energy (DOE), Office of Science, Office of Basic Energy Sciences (BES), under award number DE-FG02-06ER46327. The work at Pennsylvania State University was supported by the DOE–BES, under award number DE-SC0012375 (Y.Y., V.G.). The work at Northwestern University was supported by the Army Research Office under award numbers W911NF-15-1-0017 (J.M.R.) and DOE–BES DE-SC0012375 (D.P.). The work at the Argonne National Laboratory is supported by the DOE–BES under contract number DE-AC-02-06CH11357 (H.Z., P.J.R., Y.C., J.W.K.). The computational work made use of the Haise and Kilrain clusters at the Navy DoD Supercomputing Resource Center under the High Performance Computing Modernization Program initiative of the US Department of Defense and NSF XSEDE (ACI-1053575). The work at Cornell University was funded by DMR-1056441 (C.J.F.).

**Author Contributions** C.J.F. and C.B.E. conceived the project. T.H.K. and C.B.E. initiated the project, and fabricated and characterized the thin-film samples. C.B.E., M.S.R., V.G. and X.Q.P. supervised the experiments. D.P. and J.M.R. formulated the model and carried out the DFT study. L.X. and X.Q.P. carried out the transmission electron microscopy studies. Y.Y. and V.G. performed the SHG optical measurements. T.H.K., N.C., J.I., Y.M. and M.S.R. carried out the electrical transport measurements. T.H.K., J.P.P., J.R.P., S.R. and H.Z. performed the synchrotron CTR measurements. H.Z., J.R.P., S.R. and Y.Y. performed the COBRA analyses. Y.C., J.-W.K. and P.J.R. carried out the synchrotron diffraction and spectroscopy measurements. T.H.K., D.P., J.M.R., V.G., H.Z., M.S.R. and C.B.E. prepared the manuscript. C.B.E. directed the research.

**Author Information** Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to C.B.E. (ceom@wisc.edu).

## METHODS

**Theoretical calculations.** We perform first-principles density functional non-spin-polarized calculations within Perdew–Burke–Ernzerhof revised for solids approximation[31] as implemented in the Vienna *Ab initio* Simulation Package (VASP)[32,33] with the projector augmented wave (PAW) method[34] to treat the core and valence electrons using the following electronic configurations: $4f^1 5s^2 5p^6 6s^2$ (Nd), $3d^8 4s^2$ (Ni), $2s^2 2p^4$ (O); a $15 \times 15 \times 15$ Monkhorst–Pack $k$-point mesh[35]; and a 550 eV plane wave cut-off. For structural relaxations, we relax the atomic positions (forces to be less than 0.1 meV Å$^{-1}$) and Gaussian smearing (0.02 eV width) for the Brillouin zone (BZ) integrations. For the group theoretical analysis we use AMPLIMODES[36,37] software. Owing to the origin ambiguity in polar structures, we choose the origin of the low-symmetry structure so that the arithmetic centre remains fixed when mapping the high-symmetry structure onto the low-symmetry phase.

**Sample fabrication and characterization.** $R$NiO$_3$ thin films were epitaxially fabricated on both LaAlO$_3$ (111) and (001) substrates using pulsed laser deposition (PLD) with *in situ* reflection high energy electron diffraction (RHEED) monitoring. Before the PLD film growth, as-received LaAlO$_3$ substrates were thermally treated to obtain atomically flat surfaces with a step-terrace structure in tube furnaces at 1,100 °C for 3 h. The growth temperature and oxygen partial pressure were around 550 °C and 0.15 mbar, respectively. In a cooling process after the film deposition, *in situ* post-annealing was carried out under the oxygen ambient of 1 atm. The *in situ* RHEED patterns and atomic force microscopy (AFM) topography images were acquired for as-grown $R$NiO$_3$ thin films. We also confirmed the thickness and coherence of the $R$NiO$_3$ films using X-ray diffraction, X-ray reflectometry, and reciprocal space mappings (RSMs). In detail, RSMs around the (212) and ($\bar{1}$03) Bragg peaks of pseudocubic LaAlO$_3$ substrates were measured for $R$NiO$_3$ (111) and (001) thin films, respectively. In addition, the van der Pauw geometry was used for the electrical transport measurements of the $R$NiO$_3$ samples.

Atomically smooth surfaces with a step-terrace structure were imaged using AFM on an as-grown NdNiO$_3$ (111) film (Extended Data Fig. 3a). Spot-like RHEED patterns were observed for both thermally treated LaAlO$_3$ substrates and the as-grown nickelate films with thickness less than 10 layers (~2.2 nm), indicating that the crystalline quality of the NdNiO$_3$ film is as high as that of the underlying substrate (Extended Data Fig. 3b). After the 10th layer, one of two non-specular spots, marked by a yellow open square in Extended Data Fig. 3b, became fainter and finally disappeared in thicker layers. During the thin-film deposition, time-dependent RHEED intensity oscillations were observed clearly (Extended Data Fig. 3c). In the PLD growth of NdNiO$_3$ (111) and (001) films, one RHEED oscillation corresponds to one Ni–NdO$_3$ layer ($d$-spacing in the [111] direction ~2.2 Å) and one NdO–NiO$_2$ layer ($d$-spacing in the [001] direction ~3.8 Å), respectively. In our study, NdNiO$_3$ (111) and (001) films with the thickness of 10 layers (~2.2 nm and ~3.8 nm, respectively) were used for consistency and reliability between all experimental results.

**SHG measurements.** SHG polarimetry and temperature-dependent measurements were performed in a far-field transmission geometry using an 800 nm fundamental laser beam generated by a Coherent Evolution Nd:YLF pumped Libra-HE Ti:Sapphire femtosecond laser system (<50 fs, 2 kHz). The experimental schematic is shown in Fig. 3a, where a linear polarized fundamental field is incident on the sample at a tilt angle ($\theta$) defined by the sample normal and the incident optical wave vector. The polarization direction ($\phi$) of the incident field ($E_\omega$) was rotated through a $\lambda/2$ wave plate controlled by rotational motor. The second harmonic field ($E_{2\omega}$) generated through the nonlinear optical process inside the sample was first decomposed into $p$-polarized ($I_{2\omega||}$) and $s$-polarized ($I_{2\omega\perp}$) intensity components by a polarizing beam-splitter, then spectrally filtered and finally detected by a photo-multiplier tube. For each sample, systematic tilt-scans and polar-plots were performed by either tilting a sample by $\theta$ at a fixed $\phi$ or rotating the incident polarization by $\phi$ at a fixed $\theta$. Owing to the anisotropy between the substrate crystal physics directions [$\bar{1}\bar{1}$2] and [1$\bar{1}$0] in pseudocubic notation, two sample orientations (O1 and O2) were distinguished during the measurements, as shown in Fig. 3a.

Theoretical fitting of the SHG polarimetry data was performed using an analytical model described next. Fundamental field, ($E_\omega\cos(\phi)$, $E_\omega\sin(\phi)$, 0) with respect to the laboratory coordinates ($x$, $y$, $z$), was incident onto sample at an angle $\theta$. Domain coordinates ($Z_1$, $Z_2$, $Z_3$) defined by thin-film [$\bar{1}\bar{1}$2], [1$\bar{1}$0], [111] directions, respectively, can be described by ($\theta$, $\beta$) in laboratory coordinates, as shown in Fig. 3a and Extended Data Fig. 3d. Considering refraction and transmission at the sample surface, the fundamental field $E'_{\omega,i}$ inside the sample (′) can be expressed as

$$\begin{aligned} E'_{\omega,1} &= (\cos(\theta')\cos(\beta)\cos(\phi)t_p - \sin(\beta)\sin(\phi)t_s)E_\omega \\ E'_{\omega,2} &= (\cos(\theta')\sin(\beta)\cos(\phi)t_p + \cos(\beta)\sin(\phi)t_s)E_\omega \\ E'_{\omega,3} &= -\sin(\theta')\cos(\phi)t_p E_\omega \end{aligned}$$

where $\sin(\theta') = \sin(\theta)/n$, $n$ is the refractive index, and $t_p = 2\cos(\theta)/[n\cos(\theta) + \cos(\theta')]$ and $t_s = 2\cos(\theta)/[\cos(\theta) + n\cos(\theta')]$ are Fresnel coefficients. The SHG field ($E'_{2\omega,i}$) generated inside the sample can be calculated by $E'_{2\omega,i} = d_{ijk}E'_{\omega,j}E'_{\omega,k}$, where $d_{ijk}$ are second-order nonlinear optical coefficients. In Voigt notation, $E'_{2\omega,i} = d_{ij}E'^{\,\mathrm{Voigt}}_{\omega,j}$ and the SHG $d$ matrix for the $m$ point group symmetry is

$$d = \begin{pmatrix} 0 & 0 & 0 & 0 & d_{15} & d_{16} \\ d_{21} & d_{22} & d_{23} & d_{24} & 0 & 0 \\ d_{31} & d_{32} & d_{33} & d_{34} & 0 & 0 \end{pmatrix}$$

To simplify the analysis, we ignored the index dispersion, that is, $n = n_\omega \approx n_{2\omega}$. The SHG field in the laboratory coordinates transmitted through the sample is

$$\begin{aligned} E_{2\omega||} &= E_{2\omega,x} = (\cos(\theta')\cos(\beta)E'_{2\omega,1} + \cos(\theta')\sin(\beta)E'_{2\omega,2} - \sin(\theta')E'_{2\omega,3})t'_p \\ E_{2\omega\perp} &= E_{2\omega,y} = (-\sin(\beta)E'_{2\omega,1} + \cos(\beta)E'_{2\omega,2})t'_s \end{aligned}$$

where $t'_p = 2n\cos(\theta')/[n\cos(\theta) + \cos(\theta')]$, $t'_s = 2n\cos(\theta')/[\cos(\theta) + n\cos(\theta')]$, and the subscripts || and $\perp$ refer to fields parallel and perpendicular to the incidence plane, respectively. Thus, SHG intensity from a single domain region is $I_{2\omega||} = \alpha|E_{2\omega||}|^2$ and $I_{2\omega\perp} = \alpha|E_{2\omega\perp}|^2$, where $\alpha$ is a constant. In the above equations, $\alpha$ and $E_\omega$ act as scaling factors that can be eliminated by defining effective SHG matrices as $d_{ij}^{\mathrm{eff}} = \sqrt{\alpha}E_\omega^2 d_{ij}$. For NdNiO$_3$ thin films on the LaAlO$_3$ (111) substrate, under the phase uncorrelated approximation, SHG intensity measured by experiment can be decomposed into independent contributions from three equivalent domains denoted by different $\beta$ values, as shown in Extended Data Fig. 3d. Explicitly, we have the following equations for two substrate orientations O1 and O2

$$\text{O1:}\ \begin{cases} I_{2\omega||}^{\mathrm{total}} = w_1 I_{2\omega||}(\beta=-90°) + w_2 I_{2\omega||}(\beta=30°) + (1-w_1-w_2)I_{2\omega||}(\beta=150°) \\ I_{2\omega\perp}^{\mathrm{total}} = w_1 I_{2\omega\perp}(\beta=-90°) + w_2 I_{2\omega\perp}(\beta=30°) + (1-w_1-w_2)I_{2\omega\perp}(\beta=150°) \end{cases}$$

$$\text{O2:}\ \begin{cases} I_{2\omega||}^{\mathrm{total}} = w_1 I_{2\omega||}(\beta=0°) + w_2 I_{2\omega||}(\beta=120°) + (1-w_1-w_2)I_{2\omega||}(\beta=240°) \\ I_{2\omega\perp}^{\mathrm{total}} = w_1 I_{2\omega\perp}(\beta=0°) + w_2 I_{2\omega\perp}(\beta=120°) + (1-w_1-w_2)I_{2\omega\perp}(\beta=240°) \end{cases}$$

where $w_1$ and $w_2$ are the area fractions of two of the three domain variants in the probe area. The fits reveal these factors to be approximately 1/3 each as expected from the epitaxial geometry.

**STEM measurements and analyses.** Cross-sectional samples for STEM experiments were prepared by conventional mechanical thinning, polishing and ion milling. In the final step of the ion milling, a 0.1 kV ion beam was used to remove the beam-damaged layers on the surface. STEM-annular dark field (ADF) and ABF images were acquired using a 300 kV aberration-corrected scanning transmission electron microscope (JEOL 3100-R05) equipped with a cold field-emission gun at the University of Michigan. The convergence semi-angle of the incident electron probe and the detector angles for ADF and ABF signals are 22 mrad, 59–200 mrad and 11–22 mrad, respectively. To enhance the signal-to-noise ratio of the obtained STEM images, the images were processed by averaging along the direction parallel to the interface, for example, [$\bar{1}\bar{1}$2] direction in a NdNiO$_3$/LaAlO$_3$ (111) sample. The atomic positions of $A$-site and $B$-site cations were determined by fitting the intensity peaks with two-dimensional Gaussian functions and the relative off-centre displacement of the $A$-site atoms is defined with respect to the centre of two nearest $B$-site cations. A tilting angle ($\Theta$) in a STEM-ABF image was measured by a line connecting two nearest oxygen atoms and intersecting a line connecting two nearest $B$-site cations. An error bar was also extracted by calculating the standard deviation values from the measured tilting angles of about 10 unit cells in each layer.

**Synchrotron crystal truncation rod and coherent Bragg rod analysis.** To determine precisely the full atomic structure within each unit cells of NdNiO$_3$ thin films epitaxially grown on a LaAlO$_3$ substrate (for example, all cation and oxygen positions), we performed X-ray crystal truncation rod (CTR) measurements of both (001) and (111) thin films, and analysed CTR data using a phase retrieval technique known as coherent Bragg rod analysis (COBRA), which applies an iterative process of alternatively satisfying constraints in real and reciprocal space to reconstruct the diffraction phases from measured diffraction intensities[38,39]. The CTR measurements were conducted on a six-circle diffractometer, using an X-ray energy of 16 keV at sectors 12-ID-D and 33-ID-D of the Advanced Photon Source, Argonne National Laboratory. Both beamlines have a similar total flux of $\sim 2.0 \times 10^{12}$ photons s$^{-1}$. At 33-ID-D, the X-ray beam was focused by a pair of Kirkpatrick–Baez mirrors down to a beam profile of 50 μm (vertical) × 80 μm (horizontal). A primary twin domain of the LaAlO$_3$ substrate can be selected via the small size beam. The two-dimensional scattering images of CTRs at each step in the reciprocal lattices were recorded with a pixel array area detector (Dectris PILATUS 100 K). A large group of symmetry-inequivalent CTRs (for example, specular and non-specular, see Extended Data Fig. 9) were recorded with $L_{\max} = 4.5$

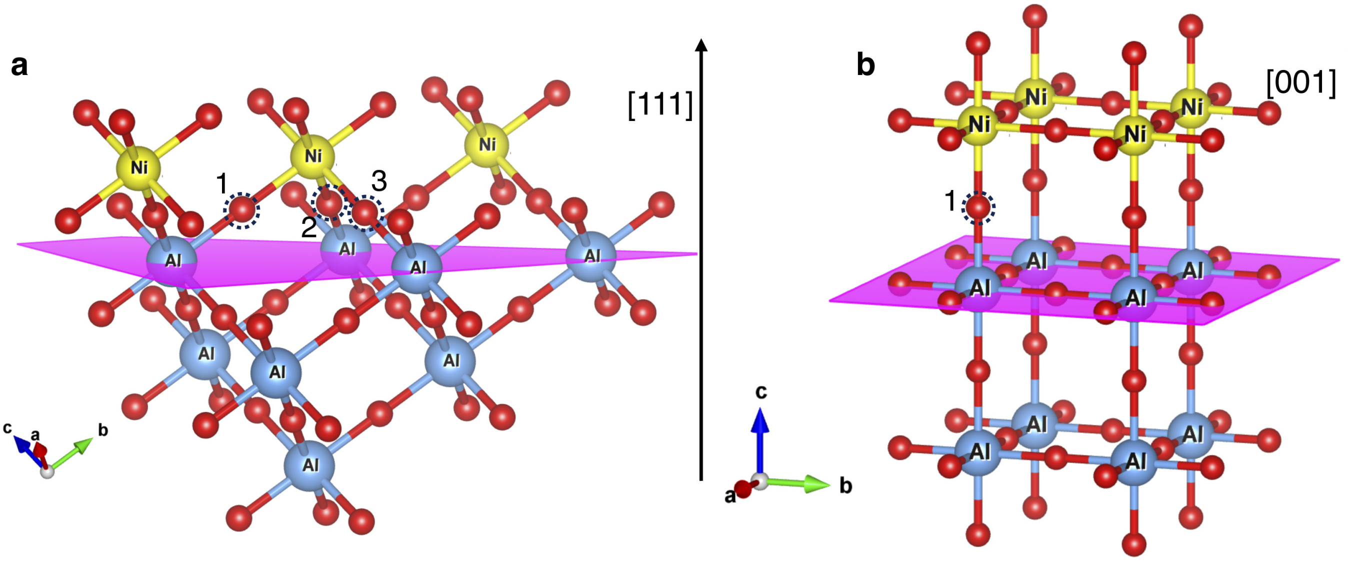
reciprocal lattice units (r.l.u.) for (001) pseudo-cubic system and $L_{max} = 15.2$ r.l.u. for (111) rhombohedral system.

The generic approach for the determination of uncertainties, based on refining a parameterized model (for example, model-dependent nonlinear least squares fitting), is not applicable to COBRA results. Instead, we adopted a method called noise analysis to estimate the error bars of the parameters of interest[40]. This method resembles the widely used bootstrap resampling approach for uncertainty estimations in statistical analysis. To estimate the error bars, we add to the experimental data random artificial noise such that the envelopes of it and its Fourier transform in real space are equal to those of the differences between the measured and COBRA-calculated CTRs. We then reanalyse the manipulated data and re-extract the parameters of interest from the density profile. This process is repeated a number of times (for example, 7–8 resamples), and the degree of scatter in the values of the parameters of interest determines their error bars.

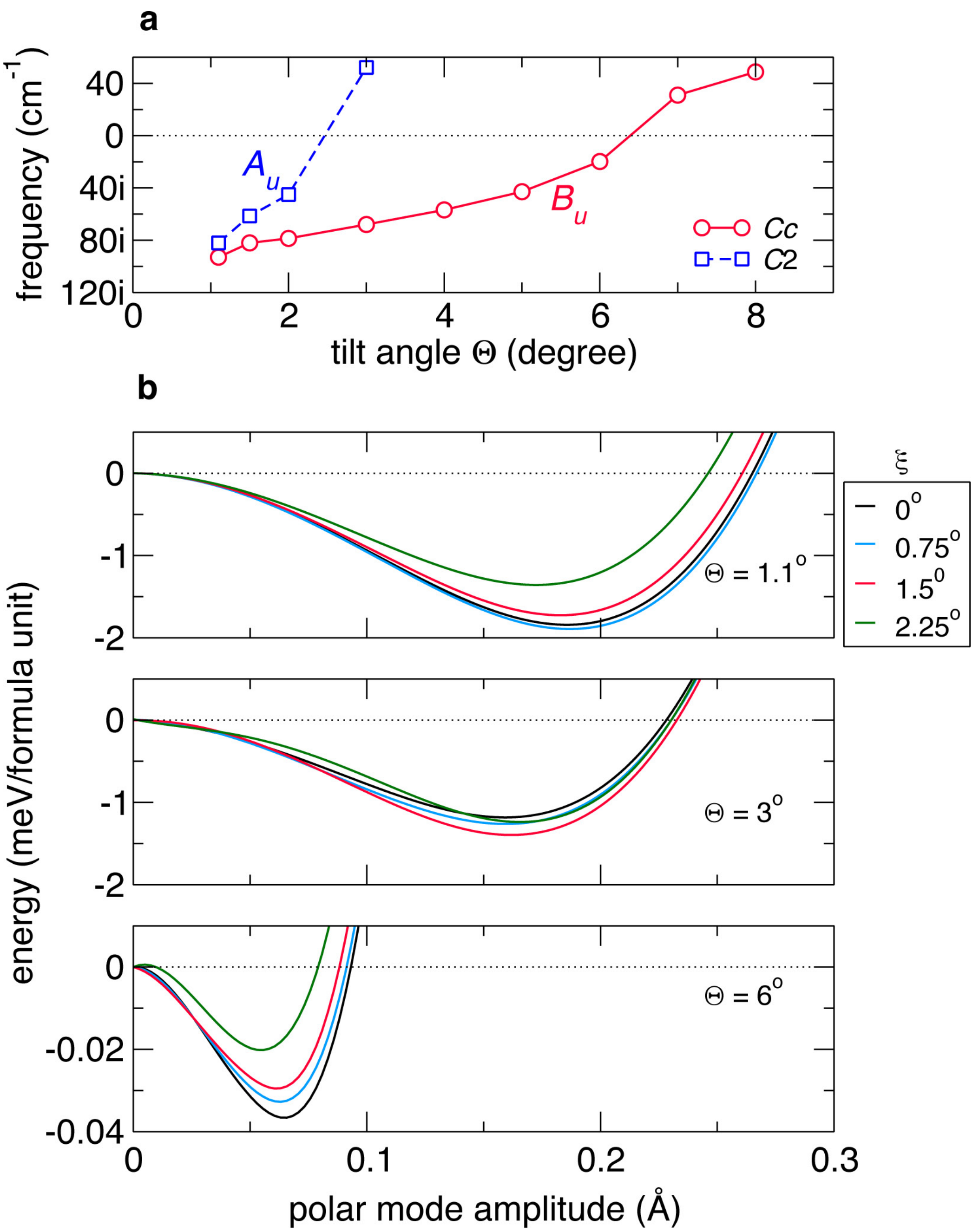
**Synchrotron X-ray diffraction and spectroscopy.** Resonant X-ray diffraction and spectroscopy experiments at the Ni K edge were performed at sectors 4-ID-D and 6-ID-B of the Advanced Photon Source, Argonne National Laboratory. X-ray absorption data were taken by monitoring Ni Ka fluorescence with an energy dispersive detector. The X-ray diffraction data for the three nearly equivalent domain population in a $NdNiO_3/LaAlO_3$ (111) film was taken with the incident X-ray energy of 16 keV.

**Hall measurements.** Transport measurements were carried out using a four-contact van der Pauw geometry over a temperature range of 80 to 300 K. Hall measurements were conducted by sourcing a d.c. current and sweeping the magnetic field over a range of −5 to 5 kG. Even components of Hall voltages, ascribed to magnetoresistance from contact misalignment, were removed and only the odd component was used to calculate $n_{3D}$. The equation $n_{3D} = I/[(dV_H/dB)\tau q]$ was used, where $I$ is the d.c. current sourced, $V_H$ is the Hall voltage, $\tau$ the thin-film thickness, and $q$ the electron charge.

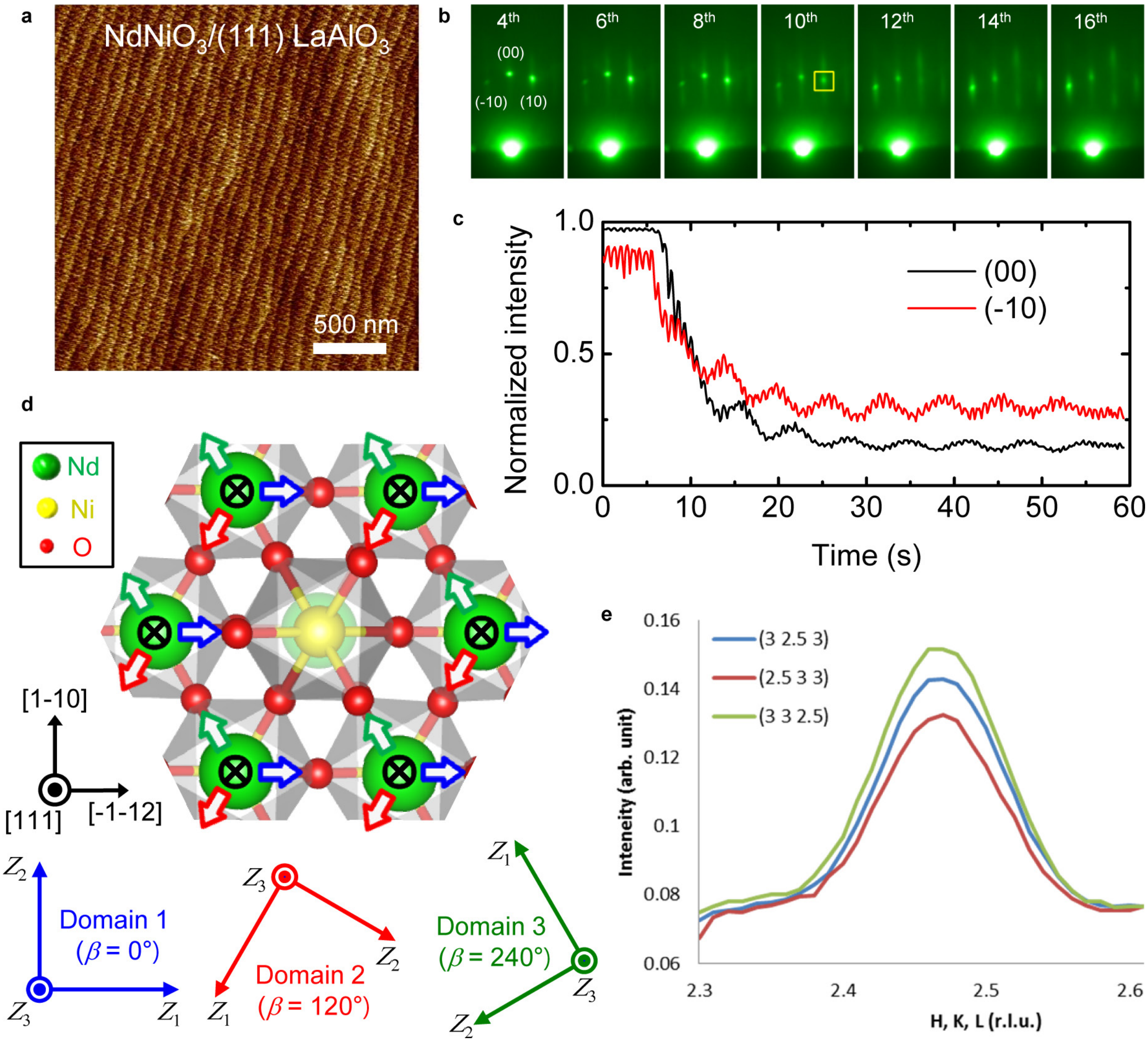

31. Perdew, J. P. *et al.* Restoring the density-gradient expansion for exchange in solids and surfaces. *Phys. Rev. Lett.* **100,** 136406 (2008).
32. Kresse, G. & Furthmüller, J. Efficient iterative schemes for *ab initio* total-energy calculations using a plane-wave basis set. *Phys. Rev. B* **54,** 11169–11186 (1996).
33. Kresse, G. & Joubert, D. From ultrasoft pseudopotentials to the projector augmented-wave method. *Phys. Rev. B* **59,** 1758–1775 (1999).
34. Blöchl, P. E. Projector augmented-wave method. *Phys. Rev. B* **50,** 17953–17979 (1994).
35. Monkhorst, H. J. & Pack, J. D. Special points for Brillouin-zone integrations. *Phys. Rev. B* **13,** 5188–5192 (1976).
36. Orobengoa, D., Capillas, C., Aroyo, M. I. & Perez-Mato, J. M. *AMPLIMODES*: symmetry-mode analysis on the Bilbao Crystallographic Server. *J. Appl. Cryst.* **42,** 820–833 (2009).
37. Perez-Mato, J. M., Orobengoa, D. & Aroyo, M. I. Mode crystallography of distorted structures. *Acta Crystallogr. A* **66,** 558–590 (2010).
38. Zhou, H. *et al.* Anomalous expansion of the copperapical oxygen distance in superconducting cuprate bilayers. *Proc. Natl Acad. Sci. USA* **107,** 8103–8107 (2010).
39. Nelson-Cheeseman, B. B. *et al.* Polar cation ordering: a route to introducing >10% bond strain into layered oxide films. *Adv. Funct. Mater.* **24,** 6884–6891 (2014).
40. Zhou, H., Pindak, R., Clarke, R., Steinberg, D. M. & Yacoby, Y. The limits of ultrahigh-resolution X-ray mapping: estimating uncertainties in thin-film and interface structures determined by phase retrieval methods. *J. Phys. D Appl. Phys.* **45,** 195302 (2012).



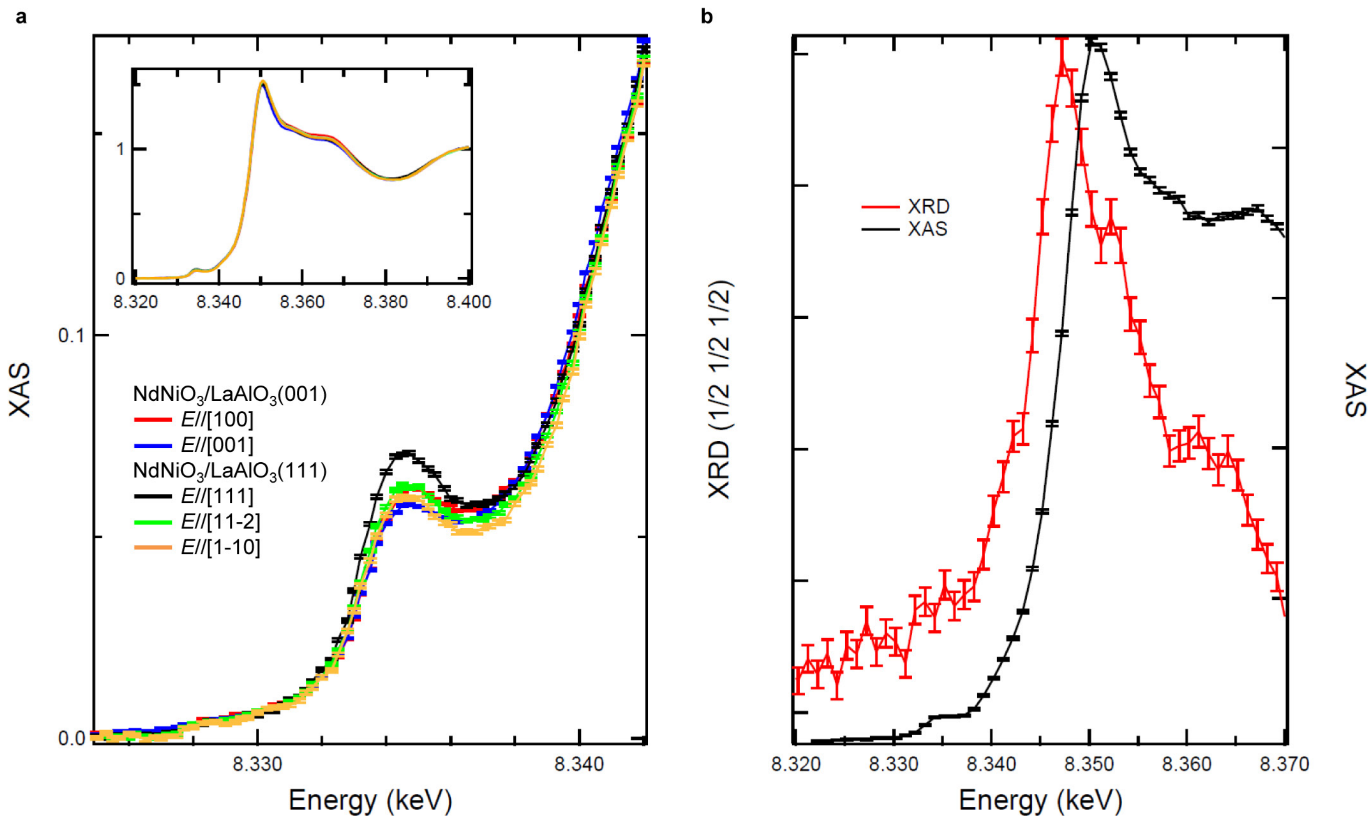

**Extended Data Figure 1 | Bond connectivity of heteroepitaxial perovskites along the (111) and (001) orientations. a**, **b**, Owing to the robust bond connectivity, the $NiO_6$ octahedra is strongly coupled with the underlying $AlO_6$ octahedra in $NdNiO_3$/$LaAlO_3$ (111) heterostructure through three bond contacts (indicated by the broken circle). The modification of oxygen octahedral structures can be better achieved in the perovskite (111) geometry (**a**) rather than the (001) geometry (**b**), which shows only one unique bond contact.



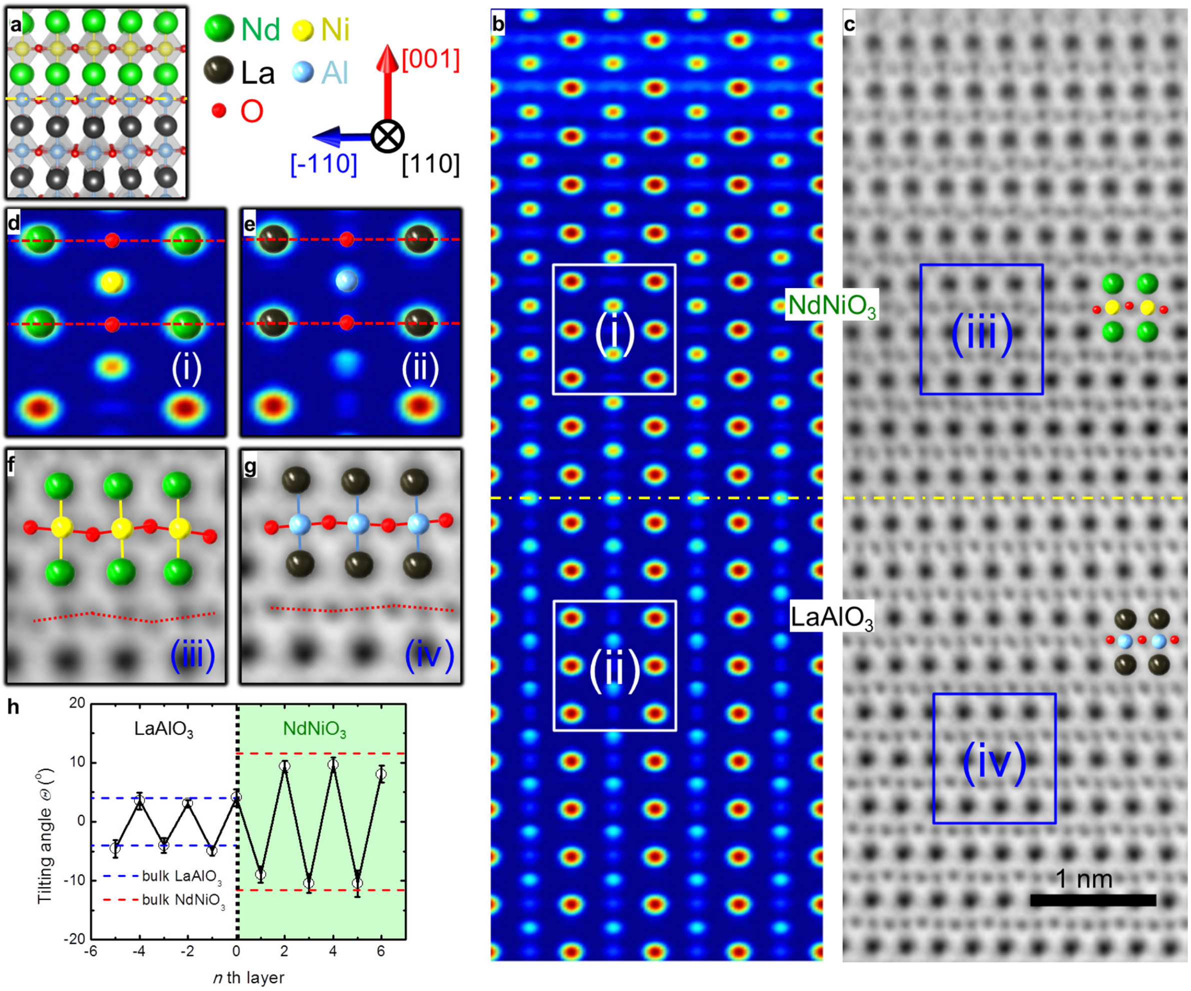

**Extended Data Figure 2 | Phonon frequencies and energy as a function of the tilting angle for $C2/c$ ($a^-a^-c^-$) $NdNiO_3$ with the (111) geometry.** **a**, At small amplitude of the tilt $\Theta$ and the out-of-phase rotation angle $\xi = 0°$ (as defined in the main text), we find two polar modes, $B_u$ and $A_u$, that lift spatial inversion symmetry resulting in the $Cc$ and $C2$ space groups, respectively (**a**). The $B_u$ and $A_u$ modes describe the polar displacements of Nd, Ni, and O atoms on the $(1\bar{1}0)_{pc}$ ($Cc$) and $(001)_{pc}$ ($C2$) planes. These two modes behave as in the $a^-a^-c^+$ case (Fig. 1b), but become hard at smaller amplitude of tilting angle, 2.45° ($A_u$) and 6.38° ($B_u$). Also, the $B_u$ mode always has a lower frequency than the $A_u$ mode. **b**, For small amplitude of the tilt ($\Theta$) and out-of-phase rotation ($\xi$) angles, we obtain an energy gain of ∼2 meV per formular unit that is smaller than the $a^-a^-c^+$ case (Fig. 1d).



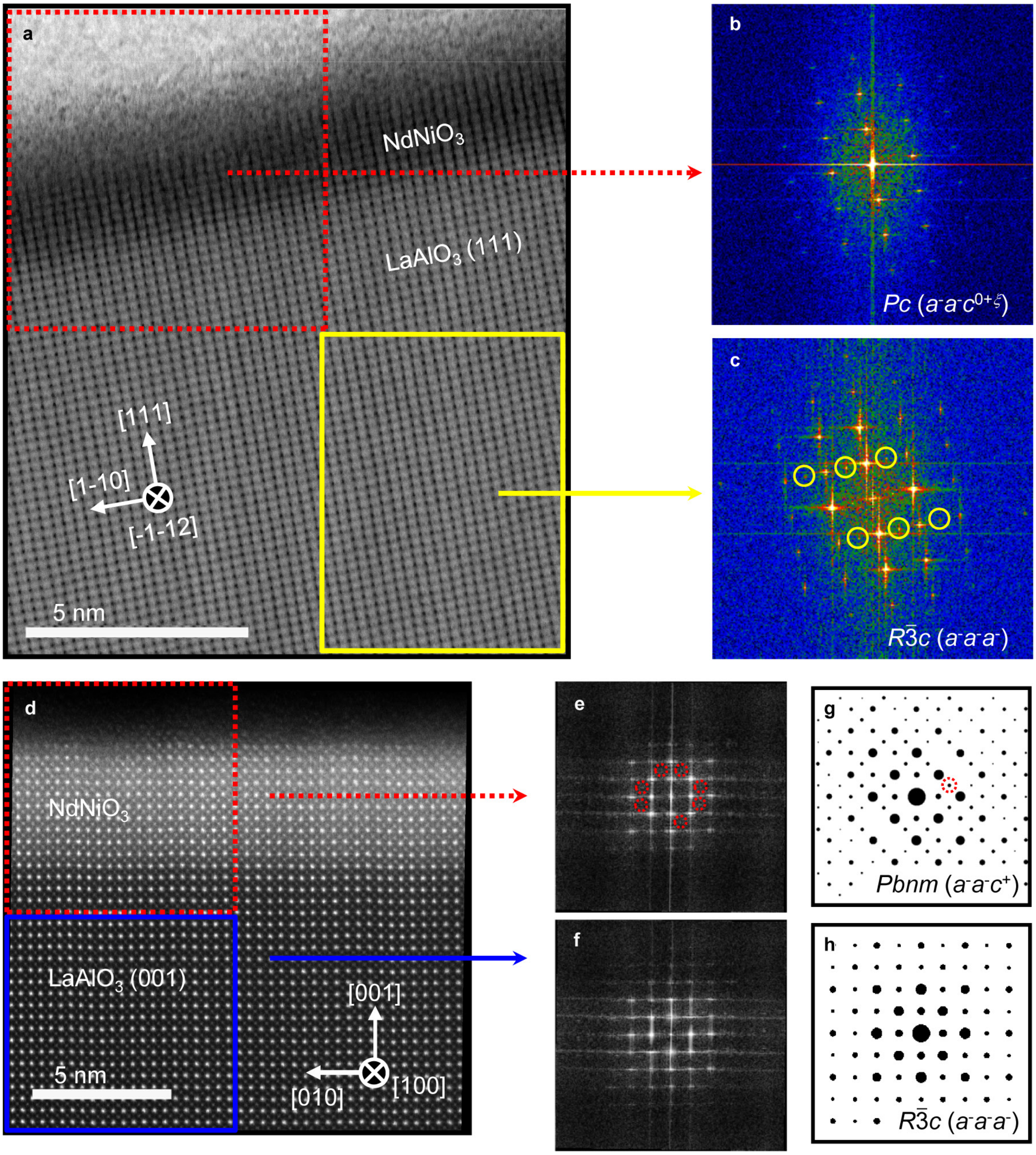

**Extended Data Figure 3 | Epitaxial synthesis and structural characterization of $NdNiO_3$/$LaAlO_3$ (111) thin films. a–c**, An atomic force microscopy (AFM) topography image (**a**) of an as-grown 2.2-nm-thick $NdNiO_3$ film, thickness-dependent evolution of *in situ* RHEED patterns (**b**), and intensity oscillation (**c**) during PLD deposition of a $NdNiO_3$ film. In **b**, a yellow square shows that a (10) RHEED peak remains up to the 10th layer, but it disappears in thicker layers. This indicates that thin $NdNiO_3$ samples with the film thickness below 10 layers (~2.2 nm) have a high crystalline quality compared with thicker $NdNiO_3$ films. In **c**, black and red lines represent time-dependent evolution of (00) and ($\bar{1}$0) peak intensity in **b**, respectively. **d**, A schematic diagram of three monoclinic domain variants, which are used for SHG polar fittings, in a $NdNiO_3$ (111) thin film. The crossed circle represents the direction of out-of-plane components of *A*-site Nd displacements. Open arrows show three possible variants of in-plane components of the *A*-site Nd displacements. Owing to the three-fold symmetry of the $LaAlO_3$ (111) substrate, three different domains exist with about equal amount. **e**, Observation of {integer integer half-integer} Bragg peaks in synchrotron X-ray diffraction. Note that the family of {integer integer half-integer} half-order Bragg peaks are directly related to the occurrence of off-symmetry *A*-site cation displacements in an orthorhombic system.



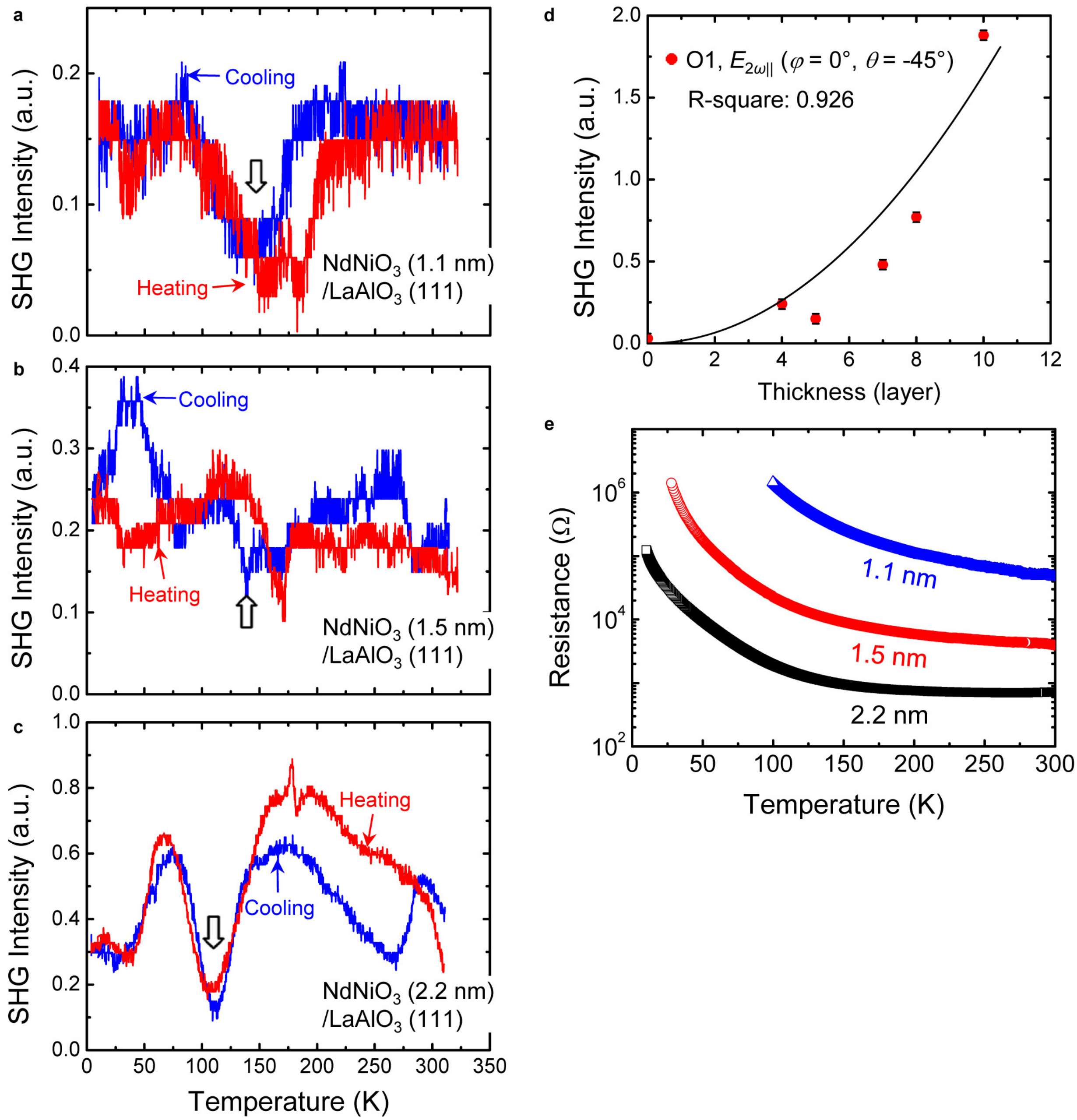

**Extended Data Figure 4 | X-ray spectroscopy and resonant X-ray diffraction measurements of epitaxial $NdNiO_3/LaAlO_3$ (111) and (001) films. a**, X-ray absorption spectroscopy (XAS) at the Ni K edge shows a clear pre-edge intensity, when the incident X-ray polarization (E) is along [111] of a $NdNiO_3$ (111) film. The near-edge X-ray absorption fine structure (NEXAFS) indicates the Ni displacement is more pronounced along $[111]_{pc}$ of the $NdNiO_3$ (111) film, while weaker along the other two in-plane directions. The response from a $NdNiO_3$ (001) film is similar to the in-plane results from the $NdNiO_3$ (111) film. The weak pre-edge intensity is due to transitions from the 1*s* to 3*d* levels. At the K edge, *s*–*d* electric dipole transition is forbidden for this octahedral case and weak quadrupole transition is allowed. However, as the central Ni atom is displaced from its cubic symmetric site, the displacement breaks the inversion symmetry, mixing *p*-state symmetry with unfilled *d*-states and allowing the dipole transition. The strength of the dipole transition is proportional to the square of the displacement along the incident X-ray polarization. **b**, Resonant X-ray diffraction (XRD) intensity at a $\left(\frac{1}{2}\frac{1}{2}\frac{1}{2}\right)$ film peak in pseudo-cubic notation (that is, equivalent to a (011) Bragg peak in orthorhombic notation) across the Ni K edge is shown. The Ni response at the $\left(\frac{1}{2}\frac{1}{2}\frac{1}{2}\right)$ arises from a combination of the finite size effect of the film (thickness fringes) and monoclinic distortion related to the Ni displacement. The error bars for both the XAS and resonant XRD results are calculated as the square root of the measured intensity.



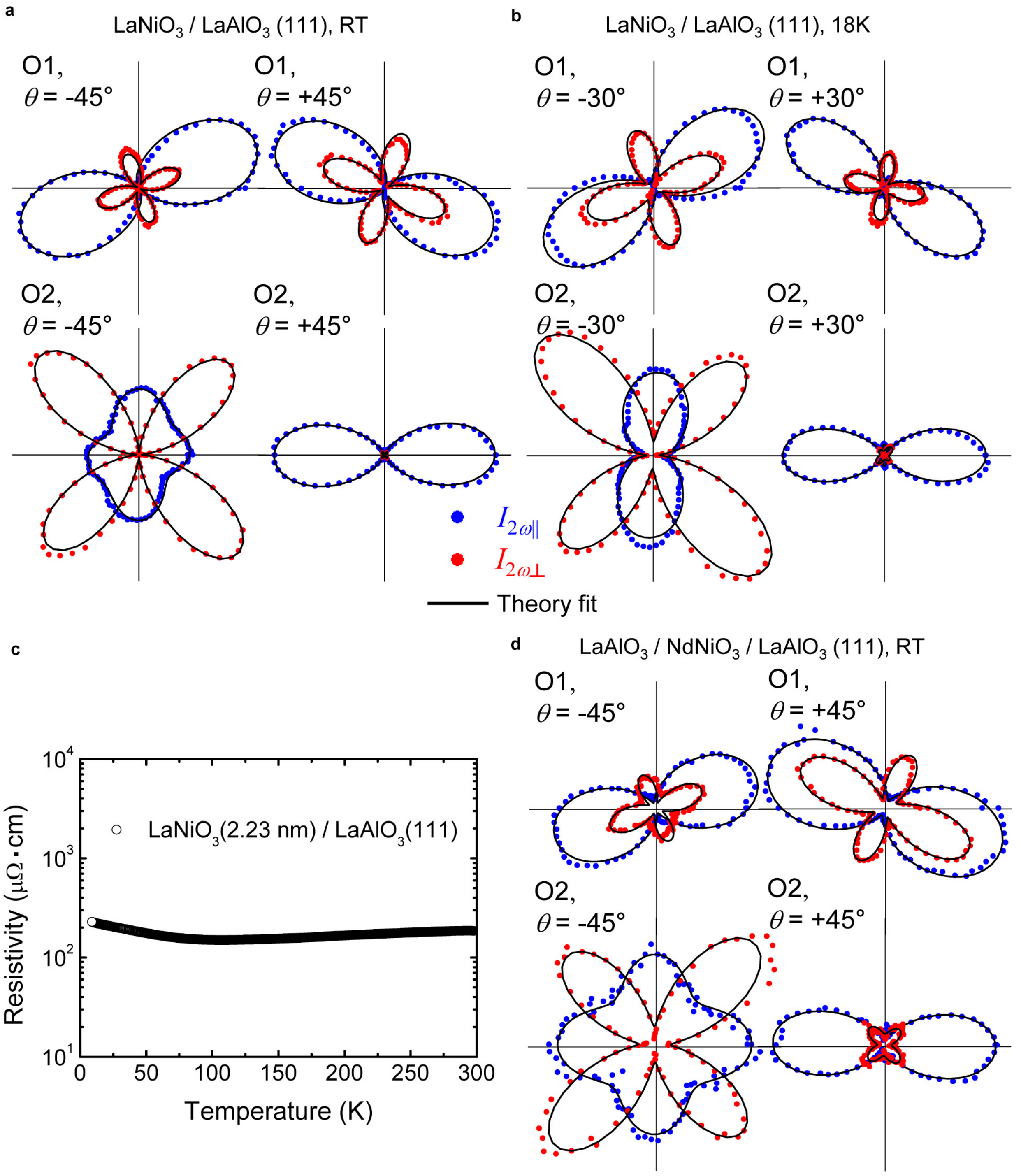

**Extended Data Figure 5 | COBRA and STEM analyses of $NdNiO_3$ thin films on $LaAlO_3$ (001) substrates. a–c**, A schematic diagram (**a**), two-dimensional electron density maps sliced through the pseudocubic (110) plane (**b**), and STEM-ABF images (**c**) of $NdNiO_3/LaAlO_3$ (001) thin-film heterostructures along the pseudocubic [110] zone axis. The interface is marked by yellow dash–dotted lines. **d**, **e**, Magnified two-dimensional electron density maps of a $NdNiO_3$ film (**d**) and a $LaAlO_3$ substrate (**e**), which are indicated by (i) and (ii) in **b**, respectively. Red dashed lines represent the positions of oxygen atoms (marked with red colours), which are taken as references to measure the relative off-centre displacements of Nd (green) and La (black) atoms, respectively. No polar displacements of the Nd and La atoms are measured with respect to the oxygen atom positions. **f**, **g**, Magnified STEM-ABF images of a $NdNiO_3$ film (**f**) and a $LaAlO_3$ substrate (**g**), which are indicated by (iii) and (iv) in **c**, respectively. Red dotted lines are guidelines to show the tilting of $NiO_6$ and $AlO_6$ octahedra in the $NdNiO_3$ and $LaAlO_3$ layers, respectively. **h**, Layer-dependent evolution of the $NiO_6$ and $AlO_6$ octahedral tilt angles across the interface in the STEM-ABF image of **c**. The 0th layer represents the $NdNiO_3/LaAlO_3$ interfaces. We extract the error bars by calculating the standard deviation values from the measured tilting angles of about 10 unit cells in each layer. Blue and red dashed lines represent the octahedral tilting angles of bulk $LaAlO_3$ and $NdNiO_3$, respectively. Note that $NiO_6$ octahedra in $NdNiO_3$ (001) thin films exhibit bulk-like tilting angle magnitudes without suppression of octahedral tilt distortion.



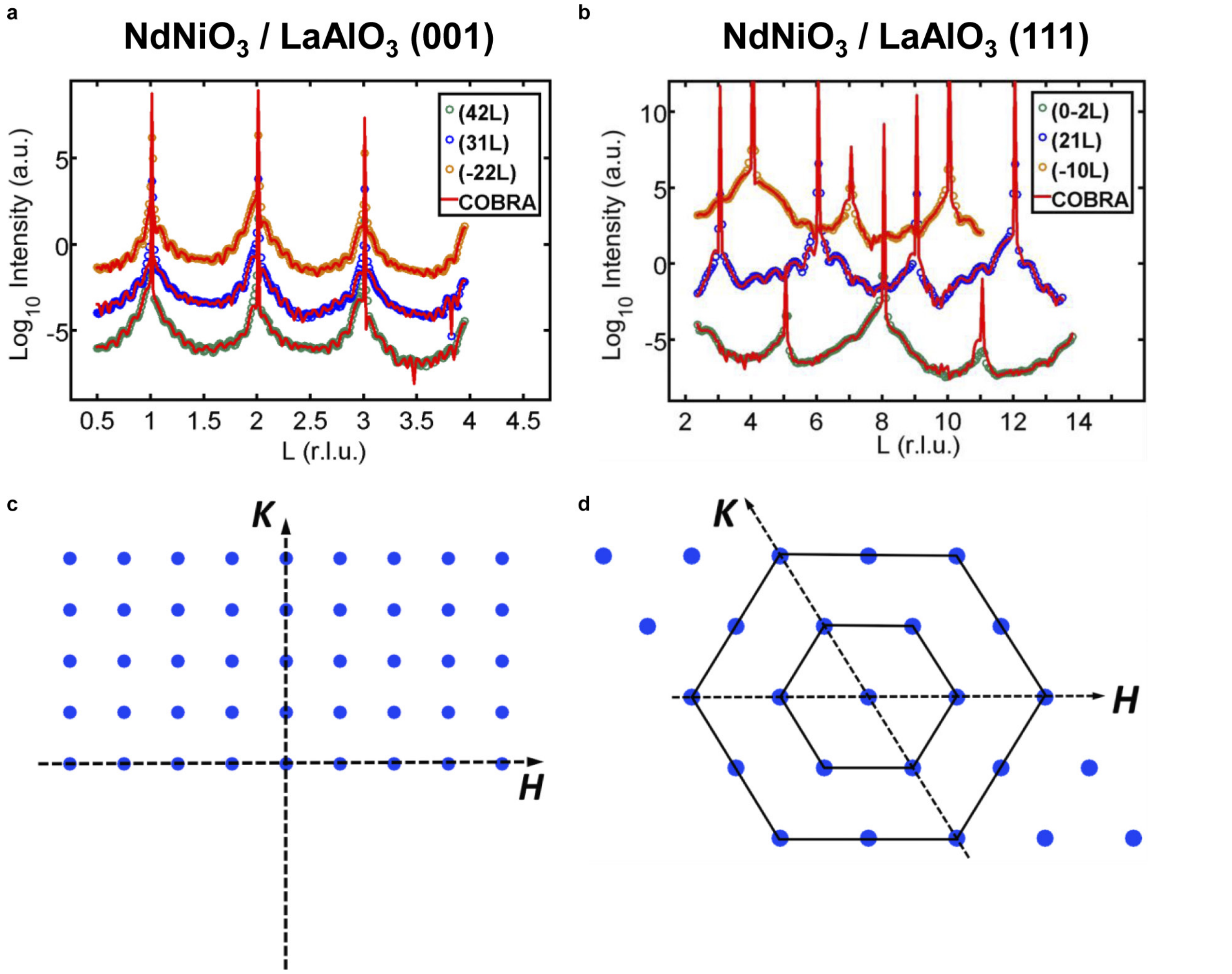

**Extended Data Figure 6 | STEM analyses of octahedral tilt patterns in $NdNiO_3/LaAlO_3$ thin films. a**, A STEM image of a $NdNiO_3/LaAlO_3$ (111) heterostructure along the pseudocubic $[\bar{1}\bar{1}2]$ zone axis. Dotted red and solid yellow squares represent the $NdNiO_3$ film and $LaAlO_3$ substrate regions for fast Fourier transform (FFT) analyses, respectively. **b**, **c**, The corresponding FFT images of the $NdNiO_3$ film (**b**) and the $LaAlO_3$ substrate (**c**) regions in **a**. In **c**, yellow circles represent half-order spots due to the $a^-a^-a^-$ tilt pattern of the $LaAlO_3$ substrate. In the FFT image of the $NdNiO_3$ film, half-order spots do not appear, indicative of local suppression of in-phase $c^+$ octahedral rotation. **d**, A STEM image of a $NdNiO_3/LaAlO_3$ (001) heterostructure along the pseudocubic [100] zone axis. **e**, **f**, The FFT images of the $NdNiO_3$ film (**e**) and the $LaAlO_3$ substrate (**f**) regions in **d**. In **e**, the dotted red circles represent half-order spots, which usually come from oxygen octahedral rotation. **g**, **h**, Simulated electron diffraction patterns of orthorhombic (*Pbnm*, $a^-a^-c^+$) $NdNiO_3$ (**g**) and rhombohedral ($R\bar{3}c$, $a^-a^-a^-$) $LaAlO_3$ (**h**) along the pseudocubic [100] zone axis. In **g**, a dotted red circle represents a half-order peak induced by in-phase $c^+$ octahedral rotation in orthorhombic $NdNiO_3$. In rhombohedral $LaAlO_3$, destructive interference occurs and then the half-order peak disappears in **h**.




**Extended Data Figure 7 | Thickness-dependent SHG and electrical transport experiments in $NdNiO_3$ (111) thin films. a–c**, The temperature-dependent SHG experiments in 1.1- (**a**), 1.5- (**b**) and 2.2-nm-thick (**c**) $NdNiO_3$ (111) thin films. **d**, The thickness dependence of the SHG responses. The SHG intensity is proportional to the square of the film thickness. **e**, The temperature-dependent resistance in 1.1-, 1.5- and 2.2-nm-thick $NdNiO_3$ (111) films.




**Extended Data Figure 8 | SHG polarimetry of LaNiO$_3$ (111) films and a capping-layer effect in NdNiO$_3$ (111) films. a**, **b**, SHG polar plots of 2.23-nm-thick LaNiO$_3$ thin films on LaAlO$_3$ (111) substrates at a room temperature (RT; **a**) and 18 K (**b**). Two SHG components are measured, $I_{2\omega||}$ (blue circles) and $I_{2\omega\perp}$ (red circles) under the O1 and O2 sample orientation, depicted in Fig. 3a. The solid lines represent the theoretical fittings of monoclinic ($m$) point-group symmetry with equivalent three domain variants. **c**, The temperature-dependent resistivity of a 2.23-nm-thick LaNiO$_3$ thin film on a LaAlO$_3$ (111) substrate. **d**, Room-temperature SHG polar plots in a NdNiO$_3$ (111) thin film with a LaAlO$_3$ capping layer.




**Extended Data Figure 9 | Synchrotron CTR measurements in $NdNiO_3$/$LaAlO_3$ thin films. a**, **b**, CTR scans of processed raw (circles) and simulated (red curves) data along various non-specular CTRs in $NdNiO_3$ (001) (**a**) and (111) (**b**) thin film, respectively. The data curves are offset for clarity of comparison. **c**, **d**, Schematics of symmetry inequivalent CTRs (*HKL*) measured in reciprocal lattices defined by the $LaAlO_3$ substrate for both $NdNiO_3$/$LaAlO_3$ (001) (**c**) and (111) (**d**) thin films. The monoclinic structure of epitaxial $NdNiO_3$ thin films breaks the high-order symmetry of the $LaAlO_3$ substrate along both (001) and (111) orientations.

**Extended Data Table 1 | Theoretical $NdNiO_3$ structure metastability for symmetry-unique structures with (111) and (001) film orientations**

(111)-Geometric Constraint

| Initial Structure | Tilt Pattern | Relaxed Structure | Energy Difference (meV/f.u.) |
|---|---|---|---|
| $P2_1/c$ | $a^-a^-c^+$ | $P2_1/c$ | 0.0 |
| $C2/m$ | $a^-a^-c^0$ | $C2/m$ | 30.3 |
| $Cm$ | $a^-a^-c^0$ | $C2/m$ | 30.3 |
| $C2/c$ | $a^-a^-c^-$ | $C2/c$ | 43.6 |
| $Cc$ | $a^-a^-c^-$ | $C2/c$ | 43.6 |
| $R\bar{3}c$ | $a^-a^-a^-$ | $R\bar{3}c$ | 85.1 |
| $R3c$ | $a^-a^-a^-$ | $R\bar{3}c$ | 85.1 |
| $R3m$ | $a^0a^0a^0$ | $R3m$ | 291 |
| $R\bar{3}m$ | $a^0a^0c^0$ | $R\bar{3}m$ | 295 |

(001)-Geometric Constraint

| | | | |
|---|---|---|---|
| $Pbnm$ | $a^-a^-c^+$ | $Pbnm$ | 0.0 |
| $C2/m$ | $a^-a^-c^0$ | $Imma$ | 36.3 |
| $Cm$ | $a^-a^-c^0$ | $Imma$ | 36.3 |
| $C2/c$ | $a^-a^-c^-$ | $C2/c$ | 49.3 |
| $Cc$ | $a^-a^-c^-$ | $C2/c$ | 49.3 |
| $Cm$ | $a^0a^0a^0$ | $Cm$ | 290 |
| $C2/m$ | $a^0a^0a^0$ | $C2/m$ | 296 |

Results from density functional theory calculations comparing stability of low-energy phases with various octahedral tilt patterns. Note that the tilt pattern does not change during the relaxation. Experimental lattice constants are used in all calculations.

doi:10.1038/nature17439

# Machine-learning-assisted materials discovery using failed experiments

Paul Raccuglia[1], Katherine C. Elbert[1], Philip D. F. Adler[1], Casey Falk[1], Malia B. Wenny[1], Aurelio Mollo[1], Matthias Zeller[2], Sorelle A. Friedler[1], Joshua Schrier[1] & Alexander J. Norquist[1]

**Inorganic–organic hybrid materials[1–3] such as organically templated metal oxides[1], metal–organic frameworks (MOFs)[2] and organohalide perovskites[4] have been studied for decades, and hydrothermal and (non-aqueous) solvothermal syntheses have produced thousands of new materials that collectively contain nearly all the metals in the periodic table[5–9]. Nevertheless, the formation of these compounds is not fully understood, and development of new compounds relies primarily on exploratory syntheses. Simulation- and data-driven approaches (promoted by efforts such as the Materials Genome Initiative[10]) provide an alternative to experimental trial-and-error. Three major strategies are: simulation-based predictions of physical properties (for example, charge mobility[11], photovoltaic properties[12], gas adsorption capacity[13] or lithium-ion intercalation[14]) to identify promising target candidates for synthetic efforts[11,15]; determination of the structure–property relationship from large bodies of experimental data[16,17], enabled by integration with high-throughput synthesis and measurement tools[18]; and clustering on the basis of similar crystallographic structure (for example, zeolite structure classification[19,20] or gas adsorption properties[21]). Here we demonstrate an alternative approach that uses machine-learning algorithms trained on reaction data to predict reaction outcomes for the crystallization of templated vanadium selenites. We used information on 'dark' reactions—failed or unsuccessful hydrothermal syntheses—collected from archived laboratory notebooks from our laboratory, and added physicochemical property descriptions to the raw notebook information using cheminformatics techniques. We used the resulting data to train a machine-learning model to predict reaction success. When carrying out hydrothermal synthesis experiments using previously untested, commercially available organic building blocks, our machine-learning model outperformed traditional human strategies, and successfully predicted conditions for new organically templated inorganic product formation with a success rate of 89 per cent. Inverting the machine-learning model reveals new hypotheses regarding the conditions for successful product formation.**

First-principles crystal-structure prediction—even for simple crystallization from a solvent—is fundamentally difficult, owing to the need to consider a combinatorially enormous set of component arrangements[22,23] using high-level quantum chemistry methods[24]. Predicting crystal structures following a chemical reaction—as in the case of hydrothermal and solvothermal synthesis—is even more challenging, because it requires an accurate potential-energy surface for the entire reaction. Instead we pose the potentially tractable question of whether a given set of reaction conditions and reagents will yield any crystal at all. A machine-learning approach to the related problem of whether a particular organic molecule will crystallize has been described previously[25]. Chemists typically posit an 'intuition' about patterns of reagent properties and composition ratios that govern material synthesis. If these patterns exist, then they can be discovered using data-mining techniques, given a database of successful and failed reactions. However, the published literature contains only a limited subset of successful reactions, typically a single set of conditions for each compound. The vast majority of unreported 'dark' (failed) reactions are archived in laboratory notebooks that are generally inaccessible. These reactions contain the valuable information needed to determine the boundaries between success and failure.

To use these data to guide future materials syntheses, we developed a web-accessible public database (http://darkreactions.haverford.edu) to facilitate both initial data entry from existing laboratory notebooks and ongoing experimental data collection. The database schema is sufficiently general to accommodate reaction descriptions beyond our particular chemical interests (for example, allowing for arbitrary numbers of inorganic and organic species, or non-aqueous solvents). We intentionally captured experimental data that might be useful for later studies (for example, product purity labels) to avoid having to re-enter experimental data, even though they were not used in the present study. The data-capture process and reliability testing are described in Methods. After excluding reactions with incomplete laboratory notebook entries, 3,955 unique, complete reactions remained for use in training and testing the machine-learning model.

Reactant names can be used to create property descriptors for our machine-learning model. For organic and oxalate-like reactants, commercially available cheminformatics software was used to compute physicochemical properties of the molecules (for example, molecular weight, number of hydrogen-bond donors/acceptors as a function of pH and polar surface area). For inorganic reactants, tabulated values of atomic properties (for example, ionization potential, electron affinity, electronegativity, hardness and atomic radius) and position on the periodic table were used. Additionally, experimental reaction conditions (for example, temperature, reaction duration and pH) and mole ratios of the different reactants were used (see Methods). A support vector machine (SVM) model was built using this expanded table of reactant properties (see Methods). The single SVM model used to predict experimental results had an accuracy of 78% in describing all of the reaction types in its test-set data, and 79% considering only vanadium-selenite reactions.

Solid-state synthesis projects can be divided into exploration and exploitation stages. Successful exploration reactions reveal new 'islands of stability'—sets of reaction conditions that result in product formation. Success rates during this stage tend to be low, because the general ranges of acceptable parameters needed for successful syntheses are unknown. The boundaries of the island can be mapped by changing the organic reactants. These exploitation reactions expand the range of functional material properties and reveal new insights about organic–inorganic interactions. Success rates during this stage can be high, because the structures and reactivities of the organic molecules can be quite similar, and so changing the organic reactants has a more subtle effect on the chemistry.

A successful model should both increase the rate of synthesis and characterization of new materials and give chemical insight.

[1]Haverford College, 370 Lancaster Avenue, Haverford, Pennsylvania 19041, USA. [2]Department of Chemistry, Purdue University, 560 Oval Drive, West Lafayette, Indiana 47907-2084, USA.

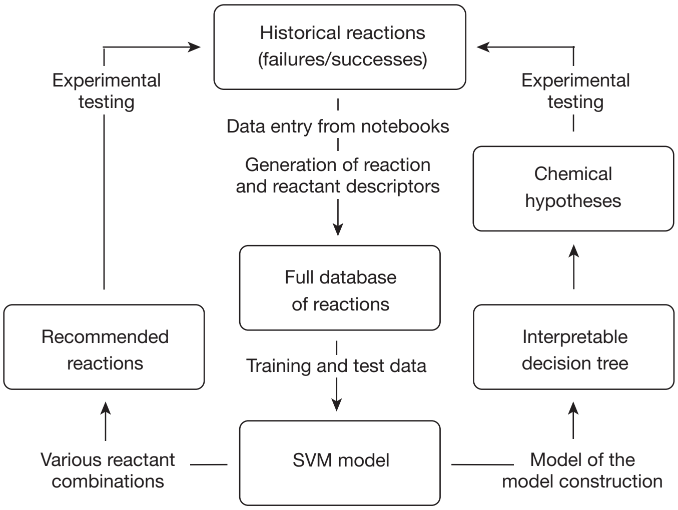



**Figure 1 | Schematic representation of the feedback mechanism in the dark reactions project.** Machine-learning models generated from historical reaction data are used to recommend new reactions to perform, and to generate human-interpretable hypotheses about crystal formation. SVM, support vector machine.

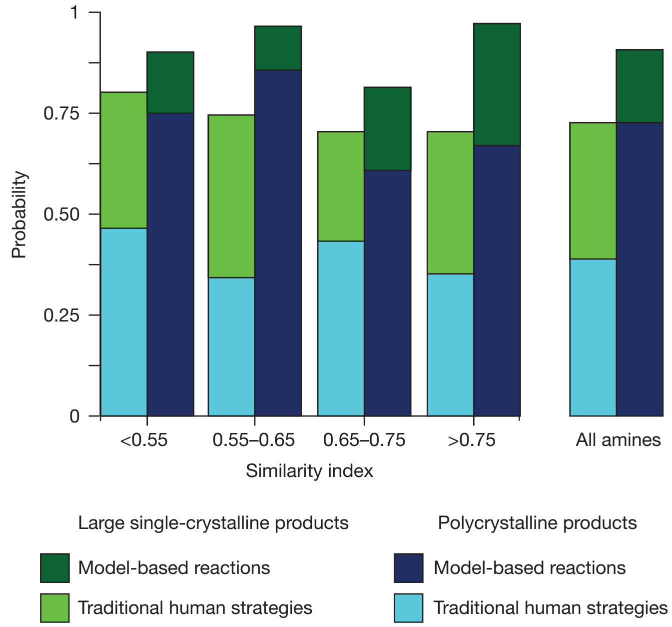



**Figure 2 | Comparison of experimental outcomes relating to the formation of templated vanadium-selenite crystals, as a function of amine similarity.** Darker coloured bars indicate model predictions; lighter colours bars indicate traditional human strategies. Reactions that yielded polycrystalline and large single-crystalline products are shown in blues and greens, respectively. The vertical axis shows the probably that the reaction had the indicated outcome. The model more successfully predicts conditions for crystal formation than do human strategies, regardless of structural similarity of the templating amines to known examples in the database.

To demonstrate the performance of our model relative to typical strategies of human chemists, we focused on exploitation reactions in templated vanadium selenites, in which a new organic building unit is introduced into a reaction. These reactions allow us to: (i) compare against the experimental decisions of experienced chemists; (ii) obtain higher quality statistical data because exploitation reactions are generally more successful; and (iii) increase understanding about the unusual degree of diversity in connectivity and dimensionality that is observed in these compounds. Though, beyond the scope of this Letter, our model could also be applied to exploration reactions, by computationally sampling possible reaction conditions involving all possible combinations of reactants, predicting successes, and then sorting the reactions by chemical interest. We used a database of commercially available organic compounds to identify 34 new diamines, sampled by structural similarity to the organic reactants already in our database (see Methods). Organically templated metal oxides using these diamines are essentially unknown, as indicated by their near absence from the Cambridge Structural Database[26] (see Methods). These amines were then used to perform human- or model-controlled hydrothermal synthesis reactions (see Methods). A schematic of this approach is shown in Fig. 1.

Reactions recommended by the model had an 89% success rate, as defined by the synthesis of the target compound type in either a polycrystalline or single-crystal form, and success rate was independent of the structural similarity of the amine (see Fig. 2). This exceeds the human intuition success rate of 78%. The difference is statistically sound. Fisher's exact test indicates better-than-chance results for model predictions with $P < 0.01$, and a two-sample proportion test indicates an 8% advantage of the model over human intuition with $P < 0.05$. The 89% success rate of the model in the experimental test is greater than the test-set accuracy measured during model construction, because the train/test split on the historical data essentially tests only exploration reactions (for which the model uncertainty is higher), whereas these experiments test exploitation reactions (for which the model uncertainty is lower).

SVMs are opaque to simple examination. To gain insight we made a 'model of the model' by re-interpreting the original SVM as a decision tree of human-interpretable if–then criteria (see Methods). An abbreviated flow-chart representation is shown in Fig. 3, and a full version of the vanadium-selenite branch of the tree is shown in Supplementary Information. From this flow chart, one can generate chemical hypotheses to guide future experiments. This approach can be applied to any chemical system for which any model exists. Here it yielded three hypotheses about the formation of templated vanadium selenites, categorized by the molecular polarizability of the amine. Representative structures for each hypothesis are shown in Fig. 4. (The model separates inorganic building units by mean Pauling electronegativity; as a consequence, vanadium selenites and molybdates appear in the same subtree. In the discussion below, we consider only the vanadium-selenite reactions contained in the subtree.)

Amines with moderate polarizability (10.29–19.51 Å$^3$), shown in blue in Fig. 3, require inclusion of a sulfur-containing reactant, specifically here V(IV)$OSO_4$. (The decision tree incidentally selects these amines by polarizability in the right branch and organic refractivity, that is, molar polarizability, in the left branch.) All but one of the organically templated vanadium selenites in the literature include $V^{4+}$ ions, which must be either introduced as a reagent or generated *in situ* through the concurrent oxidation of the amine and reduction of $V^{5+}$. These geometrically compact amines seem unable to generate the necessary $V^{4+}$ concentrations from $V^{5+}$ precursors over the timescale of the reaction. This triggers the formation of polycrystalline reaction products that do not contain the organic amines. Using V(IV)$OSO_4$ circumvents this inability to generate $V^{4+}$.

Amines with high polarizability (17.64–29.85 Å$^3$), shown in red in Fig. 3, are not limited by $V^{4+}$ generation, but do require oxalates for success. We hypothesize that oxalates alter the charge density on the inorganic secondary building unit, allowing these long, linear, highly charged tri- and tetramines to achieve charge density matching[3].

Amines with low polarizability (<9.32 Å$^3$), shown in green in Fig. 3, (for example, ethylenediamine, 1,3-diaminopropane, imidazole and N-methylethylenediamine) have higher p$K_a$ values than the other amines in our database and do not need pH < 3 to be in the correct protonation state. These amines generate sufficient $V^{4+}$ from $V^{5+}$ precursors, but slowly, requiring longer reaction times (>26 h). Use of

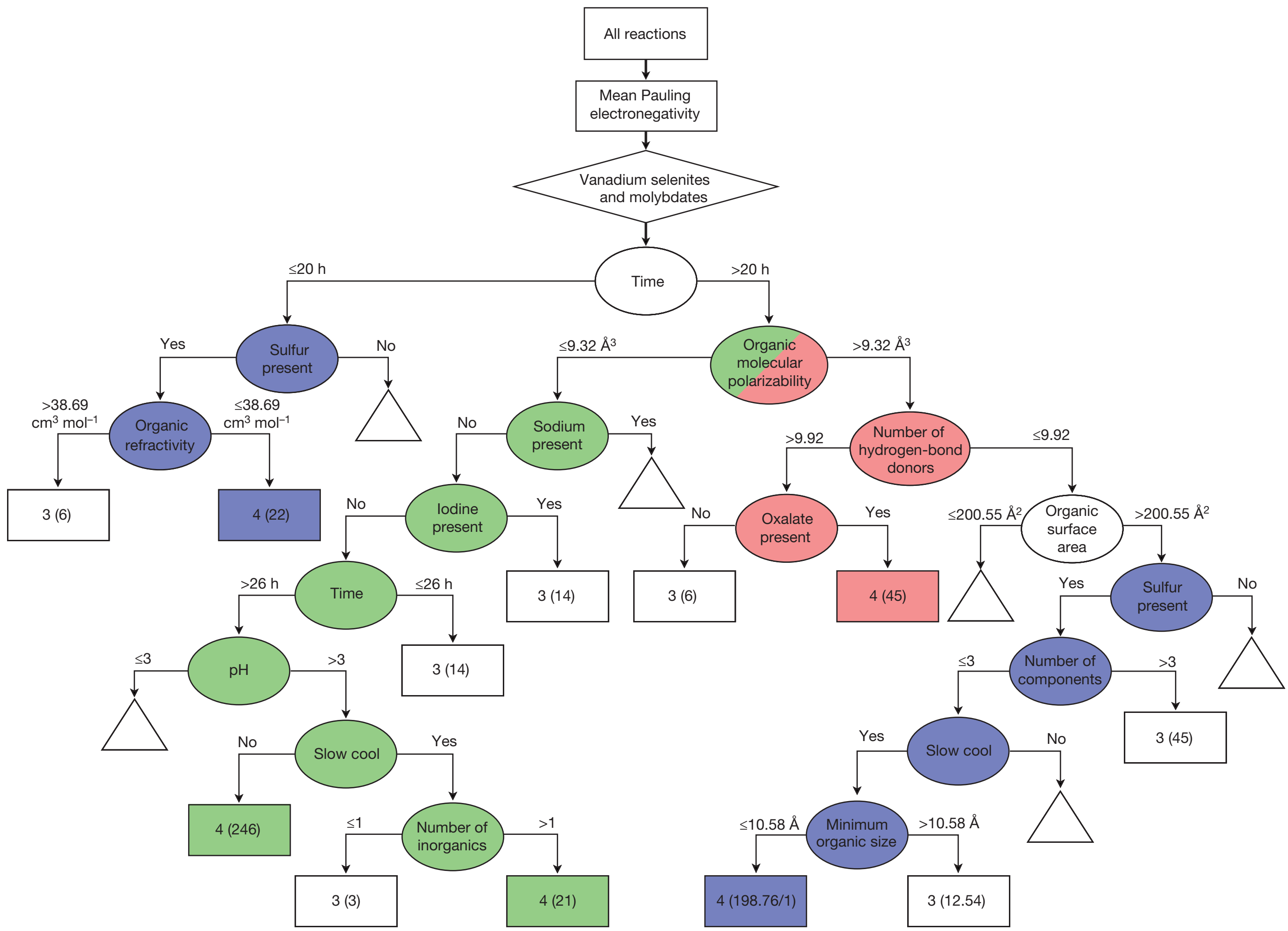



**Figure 3 | SVM-derived decision tree.** Ovals represent decision nodes, rectangles represent reaction-outcome bins and triangles represent excised subtrees. The numbers on the arrows correspond to decision attribute test values. Each reaction-outcome bin (rectangle) corresponds to a specific reaction-outcome value ('3' or '4', as indicated; see Methods); the number in parentheses is the number of reactions correctly assigned to that bin (any incorrectly classified reactions are given after a slash). Fractional values indicate reactions with an indeterminate result arising from missing attribute values higher in the tree. Bins containing the majority of successful reactions are divided into three distinct groups (indicated by green, blue and red shading). Each coloured subtree defines a specific set of reaction parameters that facilitates single-crystal formation. Inspection of these conditions leads to the corresponding chemical hypotheses, corresponding to low-, medium- and high-polarizability amines, respectively. An expanded version showing all excised subtrees is available in Supplementary Information.

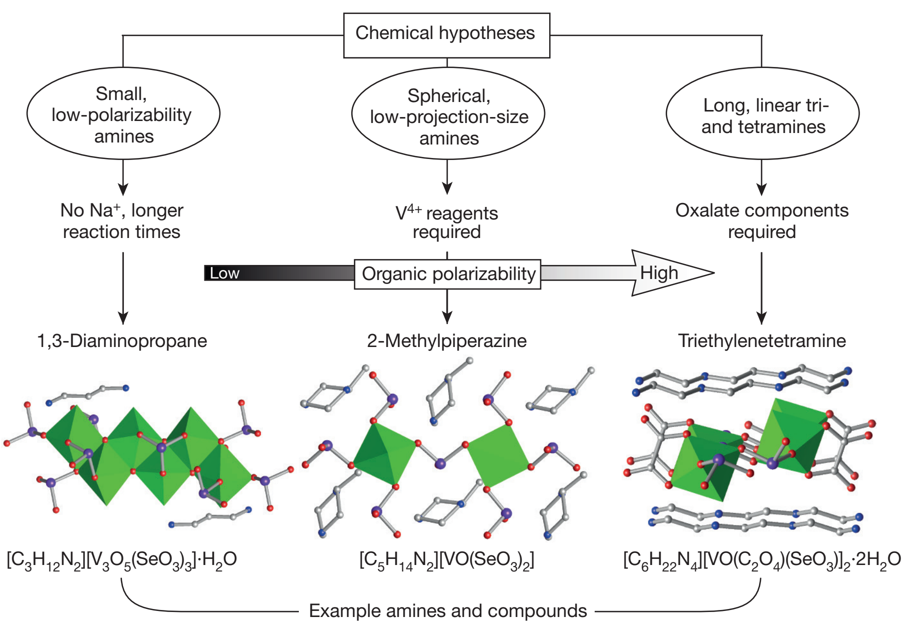



**Figure 4 | Graphical representation of the three hypotheses generated from the model, and representative structures for each hypothesis.** Experimental conditions required for single-crystal formation largely depend on the amine properties. Small, low-polarizability amines require the absence of competing $Na^+$ cations and longer reaction times, to avoid precipitating inorganic building units. Spherical, low-projection-size amines require $V^{4+}$-containing reagents such as $VOSO_4$, because they are unable to generate $V^{4+}$ directly from typical $V^{5+}$ precursors. Long tri- and tetramines require oxalate reactants, to alter the charge density of inorganic secondary building units. These three hypotheses correspond to the green, blue and red subtrees in Fig. 3, respectively.

$NaVO_3$ generally results in formation of inorganic-only polycrystalline products. Excluding sodium from the reaction mixture, by using $NH_4VO_3$, eliminates this thermodynamic sink, enabling formation of the target phase.

These hypotheses provide specific recommendations for compound formation by: (i) understanding the generation of appropriate primary building units ($V^{4+}$); (ii) enabling the construction of secondary building units that achieve charge density matching with the cationic components; and (iii) avoiding undesirable building units ($Na^+$) that result in non-templated phases. These general rules reveal previously unknown insights into our chemistry. The hypotheses derived from this analysis are manifested in three separate compounds, as shown in Fig. 4. $[C_3H_{12}N_2][V_3O_5(SeO_3)_3]\cdot H_2O$ and $[C_6H_{22}N_4][VO(C_2O_4)(SeO_3)]_2\cdot 2H_2O$ are new compounds (crystallographic details available in Supplementary Information); $[C_5H_{14}N_2][VO(SeO_3)_2]$ was reported recently[27]. The polarizabilities of the amines in these compounds range from low (1,3-diaminopropane) to moderate (2-methylpiperazine) and to high (triethylenetetramine).

Our machine-learning approach allows us to exploit chemical information contained in historical reactions and to elucidate the factors governing reaction outcome. The prediction accuracy of the model for previously untested organic amines surpassed the outcomes achieved using the chemical intuition built over many years. In addition, our approach reveals chemical principles governing reaction outcome in the form of testable hypotheses. The ability to make new compounds more successfully and to derive useful chemical information represents a transformative step forwards in exploratory reactions.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 10 September 2015; accepted 22 February 2016.**


1. Rao, C. N. R., Behera, J. N. & Dan, M. Organically-templated metal sulfates selenites and selenates. *Chem. Soc. Rev.* **35,** 375–387 (2006).
2. Zhou, H.-C., Long, J. R. & Yaghi, O. M. Introduction to metal–organic frameworks. *Chem. Rev.* **112,** 673–674 (2012).
3. Férey, G. Microporous solids: from organically templated inorganic skeletons to hybrid frameworks...ecumenism in chemistry. *Chem. Mater.* **13,** 3084–3098 (2001).
4. Stranks, S. D. & Snaith, H. J. Metal-halide perovskites for photovoltaic and light-emitting devices. *Nature Nanotechnol*. **10,** 391–402 (2015).
5. Cheetham, A. K., Férey, G. & Loiseau, T. Open-framework inorganic materials. *Angew. Chem. Int. Ed.* **38,** 3268–3292 (1999).
6. Cundy, C. S. & Cox, P. A. The hydrothermal synthesis of zeolites: history and development from the earliest days to the present time. *Chem. Rev.* **103,** 663–702 (2003).
7. Haushalter, R. C. & Mundi, L. A. Reduced molybdenum phosphates: octahedral-tetrahedral framework solids with tunnels, cages, and micropores. *Chem. Mater.* **4,** 31–48 (1992).
8. Férey, G. Oxyfluorinated microporous compounds ULM-*n*: chemical parameters structures and a proposed mechanism for their molecular tectonics. *J. Fluor. Chem.* **72,** 187–193 (1995).
9. Rao, C. N. R., Natarajan, S. & Neeraj, S. Exploration of a simple universal route to the myriad of open-framework metal phosphates. *J. Am. Chem. Soc.* **122,** 2810–2817 (2000).
10. Holdren, J. P. *et al. Material Genome Initiative Strategic Plan.* Technical Report December 2014, https://www.whitehouse.gov/sites/default/files/microsites/ostp/NSTC/mgi_strategic_plan_-_dec_2014.pdf (National Science and Technology Council, 2014).
11. Sokolov, A. N. *et al.* From computational discovery to experimental characterization of a high hole mobility organic crystal. *Nature Commun*. **2,** 437 (2011).
12. Hachmann, J. *et al.* Lead candidates for high-performance organic photovoltaics from high-throughput quantum chemistry – the Harvard Clean Energy Project. *Energy Environ. Sci.* **7,** 698–704 (2014).
13. Colón, Y. J. & Snurr, R. Q. High-throughput computational screening of metal–organic frameworks. *Chem. Soc. Rev.* **43,** 5735–5749 (2014).
14. Hautier, G., Fischer, C. C., Jain, A., Mueller, T. & Ceder, G. Finding nature's missing ternary oxide compounds using machine learning and density functional theory. *Chem. Mater.* **22,** 3762–3767 (2010).
15. Martin, R. L., Lin, L.-C., Jariwala, K., Smit, B. & Haranczyk, M. Mail-order metal–organic frameworks (MOFs): designing isoreticular MOF-5 analogues comprising commercially available organic molecules. *J. Phys. Chem. C* **117,** 12159–12167 (2013).
16. Gaultois, M. W. *et al.* Data-driven review of thermoelectric materials: performance and resource considerations. *Chem. Mater.* **25,** 2911–2920 (2013).
17. Kalidindi, S. R. & Graef, M. D. Materials data science: current status and future outlook. *Annu. Rev. Mater. Res.* **45,** 171–193 (2015).
18. Zhao, J.-C. High-throughput experimental tools for the materials genome initiative. *Chin. Sci. Bull.* **59,** 1652–1661 (2014).
19. Yang, S., Lach-hab, M., Vaisman, I. I. & Blaisten-Barojas, E. Identifying zeolite frameworks with a machine learning approach. *J. Phys. Chem. C* **113,** 21721–21725 (2009).
20. Li, Y. & Yu, J. New stories of zeolite structures: their descriptions, determinations, predictions, and evaluations. *Chem. Rev.* **114,** 7268–7316 (2014).
21. Fernandez, M., Boyd, P. G., Daff, T. D., Aghaji, M. Z. & Woo, T. K. Rapid and accurate machine learning recognition of high performing metal organic frameworks for $CO_2$ capture. *J. Phys. Chem. Lett.* **5,** 3056–3060 (2014).
22. Groom, C. R. & Reilly, A. M. Sixth blind test of organic crystal-structure prediction methods. *Acta Crystallogr.* **B70,** 776–777 (2014).
23. Thakur, T. S., Dubey, R. & Desiraju, G. R. Crystal structure and prediction. *Annu. Rev. Phys. Chem.* **66,** 21–42 (2015).
24. Beran, G. J. O. A new era for ab initio molecular crystal lattice energy prediction. *Angew. Chem. Int. Ed.* **54,** 396–398 (2015).
25. Wicker, J. G. P. & Cooper, R. I. Will it crystallise? Predicting crystallinity of molecular materials. *CrystEngComm* **17,** 1927–1934 (2015).
26. Allen, F. H. The Cambridge Structural Database: a quarter of a million crystal structures and rising. *Acta Crystallogr.* **B58,** 380–388 (2002).
27. Olshansky, J. H. *et al.* Formation principles for vanadium selenites: the role of pH on product composition. *Inorg. Chem.* **53,** 12027–12035 (2014).


**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** We thank Y. Huang, G. Martin-Noble and D. Reilley for data entry and J. H. Koffer for synthetic efforts. M.Z. acknowledges support for the purchase of a diffractometer by the National Science Foundation (DMR 1337296), the Ohio Board of Reagents grant CAP-491 and from Youngstown State University. This work was supported by the National Science Foundation (DMR-1307801). A.J.N. and J.S. each acknowledge the Henry Dreyfus Teacher-Scholar Award program.

**Author Contributions** S.A.F., J.S. and A.J.N. conceived the project and wrote the paper. A.J.N. supervised the data capture. C.F. developed the web-accessible database. A.J.N. and P.D.F.A. tested the data reliability. J.S. and P.R. developed the reactant descriptors. P.R., C.F. and S.A.F. developed the machine-learning models. J.S. performed diamine selection. P.D.F.A performed the Cambridge Structural Database search. K.C.E., M.B.W. and A.M. performed the hydrothermal experimental reactions, supervised by A.J.N. M.Z. performed X-ray crystallography on the resulting products. P.D.F.A. performed the statistical analyses. P.D.F.A., A.J.N., J.S. and S.A.F. performed the decision-tree calculation and analysis. All authors discussed the results and commented on the manuscript.

**Author Information** Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to S.A.F. (sorelle@cs.haverford.edu), J.S. (jschrier@haverford.edu) or A.J.N. (anorquis@haverford.edu).

## METHODS

**Data capture and reliability.** The average rate of data entry from our laboratory notebooks was approximately 50 reactions per hour. Three types of data were entered from the laboratory notebooks. First, compositional information was entered in the form of reactant identities and quantities. Reactants were categorized as being building units for the organic or inorganic structures, or acting as solvent (water). Second, reaction conditions were described, including initial solution pH and heating profile data. Third, reaction-outcome data included both qualitative descriptions of the products and product purity. These descriptions were coded during data entry. Crystal size was coded with the labels 1 for no solid product, 2 for an amorphous solid, 3 for a polycrystalline sample or 4 for single crystals with average crystallite dimensions exceeding approximately 0.01 mm. (This size corresponds to the general requirements for standard single-crystal X-ray diffraction data collection.) Product purity was coded with the labels 1 for a multiphase product or 2 for a single-phase product.

Reliability testing was performed on 100 randomly selected reactions from the database. Each field in each reaction was checked against the laboratory notebook from which this entry was generated. The overall error rate for all fields was 1.89%, which corresponds to 34 errors from a set of 1,800. Each reaction must have at least one inorganic component, one organic component, one solvent, as well as all reaction conditions and outcomes fields listed above. If any of these fields is missing, the reaction is entered into the database for completeness, but is not used for the training or testing of the machine-learning model described below. These filters resulted in a dataset of 3,955 unique, complete reactions.

**Reactant descriptors.** The ChemAxon Calculator Plugins[28] were used to compute the physicochemical properties of the organic and oxalate-like reactants (for example, molecular weight, number of hydrogen-bond donors/acceptors as a function of pH and polar surface area). For both the organic and oxalate-like reactants, 19 properties were used directly, and others were used to calculate 6 variables describing the mole ratios of the different reactants that were present. For inorganic reactants, 12 atomic properties (for example, ionization potential, electron affinity, electronegativity, hardness and atomic radius), 22 logical values describing the presence or absence of particular metal types, 28 logical values describing the position on the periodic table, and 8 logical values describing the metal valence were used for each element type contained in the reactants. Five variables are experimental reaction conditions (for example, temperature, reaction duration and pH). The descriptor variables are represented in a permutation-invariant fashion (maximum, minimum, arithmetic- and geometric- means) for each reactant type, so that neither the order in which the data are entered nor the number of each component matters, which results in a total of 273 descriptors per reaction. See Supplementary Information for a complete table of computed physicochemical properties.

**SVM creation and validation.** A broad set of models was evaluated, including decision trees, random forests, logistic regression, $k$-nearest neighbours and SVMs[29]. As shown in Supplementary Table 5, a SVM resulted in the highest accuracy, 74%, as measured using a calculated average of 15 training/test splits. Specifically, a SVM[30] model with a universal Pearson VII function-based kernel[31] was trained on 3,955 labelled reactions previously performed by the laboratory. The SVM was implemented in WEKA 3.7[32,33]; this implementation included a built-in data-normalization step. The model was tested against the known data for its accuracy using a standard 1/3-test and 2/3-training data split. Because the goal is to predict the outcome of reactions with new combinations of reactants, careful partitioning of the test set was required. Holding out test data uniformly at random would potentially put the same combinations of inorganic and organic reactants (reactions differing only by stoichiometries and other conditions) into both the test and training sets, and thus artificially inflate the accuracy rate. Instead, all of the reactions containing a particular set of inorganic and organic reactants were placed into either the test or training set. Under these conditions, the SVM model was measured according to its two-class accuracy, where outcomes of '3' or '4' were considered successes and '1' and '2' were grouped together as failed reactions. The single SVM model used to predict experimental results had an accuracy of 78% in describing all of the reaction types in its test-set data, and 79% considering only vanadium-selenite reactions. The average over 15 such splits was 74%. A learning curve was constructed to test the SVM; details are available in Supplementary Information.

High-dimensional feature spaces are not problematic for SVMs, because they are especially robust to correlated features and are frequently used for problems with many more dimensions than our feature set (for example, in textual learning with 10,000 features)[34]. Feature selection was performed on the model to identify the properties with the most influence on classification success (see Supplementary Information). The selected features were properties of the organic amines (van der Waals surface area, solvent-accessible surface area of positively charged atoms and the number of hydrogen-bond donors) and the inorganic components (mean of the Pauling electronegativites of the metals, their mole-weighted hardness and mean mole-weighted atomic radii). Using only these six features lowers the model accuracy to 70.7%; therefore, the entire set of features was used for the experimental tests. However, the six selected features listed above appear in the decision-tree description of the model.

**Selection of new diamines.** The eMolecules database (http://academics.emolecules.com/) was used to identify new diamines comprised of only C, H and N atoms, excluding nitriles, hydrazines and isotopically labelled compounds, resulting in 1,680 previously untested, commercially available diamines. For each diamine, a structural fingerprint based on the topological bond paths[35] of the molecule was calculated, and the maximum structural similarity to any of the existing organic compounds in the database was computed using the Tanimoto similarity[35]; the fingerprinting and similarity calculations were performed using the default parameters of the RDKit (http://www.rdkit.org). The particular similarity measure used is not crucial—a comparison of 12 standard fingerprinting methods found that they are all correlated with one another[36]. The list was ranked by similarity and by cost, using the Sigma Aldrich (338 diamines) and Alfa Aesar (62 additional diamines) catalogue prices. After excluding the highest-cost diamines, we sampled 34 diamines across the range of similarities to existing compounds. The same 34 amines were used for both the model and human reactions discussed in the text.

On average, 2 structures have been reported for each of the 34 diamines in the Cambridge Structural Database (CSD)[26], with 19 not existing in any templated metal-oxide structure in the CSD. By contrast, an average of 151 unique structures exist for the most frequently used amines (piperazine, ethylenediamine, 4,4′-dipyridyl and DABCO).

**Hydrothermal synthesis.** To avoid introducing biases, all reaction types (which differ in specific sets of reagents and reaction conditions) were randomly assigned to be human- or model-controlled, with the stipulation that each amine appear with approximately the same frequency. Amine quantities were determined by either the model or an approach that simply captures human intuitions about exploitation reactions. The recommendations of the model were generated by sampling a range of organic mole amounts, then sorting the results by predicted outcome and confidence. For consistency, human reactions used a rule-based approach that is widely used by the exploratory hydrothermal synthesis community[37], namely, scaling the masses of the organic amines by their respective formula weights, while all other reaction parameters remain unchanged. For brevity, we call this rule-based approach to capture human chemical knowledge "intuition". All reactions were conducted under mild hydrothermal conditions, in 23-ml poly(fluoroethylene-propylene)-lined pressure vessels. The pH values of the initial reaction mixtures were adjusted to the appropriate values using either 4 M HCl or 4 M NaOH. Reaction mixtures were heated to 90–110 °C for 12–72 h. Pressure vessels were opened in air after reaction and products were recovered through filtration. Objective metrics (measured crystallite size and powder X-ray diffraction) were used to score reaction outcomes.

**Statistical analysis.** Statistical analyses were performed with standard packages available in R 3.2.1[38]. No statistical methods were used to predetermine sample size.

**Decision-tree construction.** All data were relabelled with the predicted outcomes of the SVM model and a C4.5 decision tree (implemented in WEKA 3.7)[32] was used to model those predicted outcomes[39].

**Code availability.** All code for this project is available at https://github.com/darkreactions. The code is licensed under the GPL version 3. The precise terms of said license are available with the code.


28. JChem 6.1.3, http://www.chemaxon.com (ChemAxon, 2013).
29. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning* 2nd edn, Ch. 9, 12, 13, 15 (Springer, 2009).
30. Cortes, C. & Vapnik, V. Support-vector networks. *Mach. Learn.* **20,** 273–297 (1995).
31. Üstün, B., Melssen, W. & Buydens, L. M. C. Facilitating the application of support vector regression by using a universal Pearson VII function based kernel. *Chemom. Intell. Lab. Syst.* **81,** 29–40 (2006).
32. Hall, M. *et al.* The WEKA data mining software: an update. *ACM SIGKDD Explor. Newslett.* **11,** 10–18 (2009).
33. Chang, C.-C. & Lin, C.-J. LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol.* **2,** 27 (2011).
34. Joachims, T. Text categorization with support vector machines: Learning with many relevant features. In *Proc. 10th European Conf. Machine Learning* (eds Nédellec, C. & Rouveirol, C.) 137–142 (Springer, 1998).
35. Leach, A. & Gillet, V. J. *An Introduction to Chemoinformatics* Ch. 5 (Springer, 2007).
36. Riniker, S. & Landrum, G. A. Open-source platform to benchmark fingerprints for ligand-based virtual screening. *J. Cheminformat.* **5,** 26 (2013).
37. Thangavelu, S. G., Butcher, R. J. & Cahill, C. L. Role of N-donor sterics on the coordination environment and dimensionality of uranyl thiophenedicarboxylate coordination polymers. *Cryst. Growth Des.* **15,** 3481–3492 (2015).
38. R Core Team. *R: A Language and Environment for Statistical Computing* http://www.R-project.org/ (R Foundation for Statistical Computing, 2015).
39. Barakat, N. & Diederich, J. Eclectic rule-extraction from support vector machines. *Int. J. Comput. Intell.* **2,** 59–62 (2005).

doi:10.1038/nature17431

# Scalable and sustainable electrochemical allylic C–H oxidation

Evan J. Horn[1]*, Brandon R. Rosen[1]*, Yong Chen[2], Jiaze Tang[2], Ke Chen[3], Martin D. Eastgate[3] & Phil S. Baran[1]

**New methods and strategies for the direct functionalization of C–H bonds are beginning to reshape the field of retrosynthetic analysis, affecting the synthesis of natural products, medicines and materials[1]. The oxidation of allylic systems has played a prominent role in this context as possibly the most widely applied C–H functionalization, owing to the utility of enones and allylic alcohols as versatile intermediates, and their prevalence in natural and unnatural materials[2]. Allylic oxidations have featured in hundreds of syntheses, including some natural product syntheses regarded as "classics"[3]. Despite many attempts to improve the efficiency and practicality of this transformation, the majority of conditions still use highly toxic reagents (based around toxic elements such as chromium or selenium) or expensive catalysts (such as palladium or rhodium)[2]. These requirements are problematic in industrial settings; currently, no scalable and sustainable solution to allylic oxidation exists. This oxidation strategy is therefore rarely used for large-scale synthetic applications, limiting the adoption of this retrosynthetic strategy by industrial scientists. Here we describe an electrochemical C–H oxidation strategy that exhibits broad substrate scope, operational simplicity and high chemoselectivity. It uses inexpensive and readily available materials, and represents a scalable allylic C–H oxidation (demonstrated on 100 grams), enabling the adoption of this C–H oxidation strategy in large-scale industrial settings without substantial environmental impact.**

Electrochemical oxidation presents an attractive alternative to traditional chemical reagents for large-scale applications, in a large part owing to the generation of less toxic waste than that produced by current chemical processes[4,5]. In addition, electrochemical conditions are compatible with a wide range of functional groups[6–10], tend to have higher overall energy efficiency as compared to thermal processes and, owing to their limited use, offer new intellectual property space for small-molecule synthesis[4]. The first electrochemical allylic oxidation was reported in 1968[11,12] (Fig. 1c). Direct oxidation of $\alpha$-pinene (**1**) led to the fragmentation of the cyclobutane ring with incorporation of methanol or acetic acid (depending on the solvent) to give products **2** in 22%–24% yield. A major advance in this field came in 1985, when the indirect oxidation of **1** using *N*-hydroxyphthalimide (NHPI) as an electrochemical mediator was reported[13–15] (Fig. 1c). Unfortunately, verbenone (**3**) was isolated in only 13%–23% yield. Although these reactions are not useful in a preparative sense, they were a proof of concept that served as a foundation for our work.

In our own laboratory, systematic and extensive experimentation led to the identification of three modifications of the original precedent[13–15], which transformed this process into a synthetically useful electrochemical allylic C–H oxidation (Fig. 2). As described below, these modifications include the addition of a simple co-oxidant, the identification of a new electrochemical mediator, and the design of a reliable and inexpensive set-up.

From the outset of this work, we avoided the use of expensive electrodes such as precious metals (for example, platinum or gold), focusing our efforts exclusively on carbon. Initial optimization was undertaken using graphite rods, but despite clean conversion of starting material to desired product, mass recovery was typically low. We considered that this might have partially been due to absorption of the substrate onto the graphite. Switching to reticulated vitreous carbon (100 pores per inch, acquired from K. R. Reynolds Co. for about US$3 per electrode) electrodes proved to be far more productive.

In our laboratory, the original conditions[14] applied to valencene (**4**) led to only 6% isolated yield of nootkatone (**5**), the principal fragrance component of grapefruit aroma (Fig. 2). Our hypothesis was that air was the oxygen-atom source in this transformation, which was qualitatively confirmed by bubbling $O_2$ gas in the reaction, resulting in an improved isolated yield of 18%. However, NHPI/$O_2$ systems[16–18] have been explicitly avoided by the pharmaceutical industry, along with other oxygen-mediated reactions, owing to the challenges relating to flammability, and other issues arising from reliably and safety in performing oxygen-mediated reactions on a large scale[19,20]; as such, applications of aerobic oxidations in the pharmaceutical and fine-chemical industries remain sparse[21–23]. Thus, a number of co-oxidants were evaluated, using NHPI as a mediator, and *tert*-butyl hydroperoxide ($^t$BuOOH) led to substantial increases in reaction conversion and reproducibility, delivering **5** in 51% yield. Using $^t$BuOOH without the NHPI mediator led to only 18% isolated yield under otherwise identical conditions.

With a suitable co-oxidant selected, attention turned to the optimization of base, solvent and electrolyte. We evaluated a variety of organic and inorganic bases, with pyridine proving to be ideal. The use of acetone as solvent led to a slightly increased yield (56% **5** isolated) and was chosen as a general solvent for this reaction owing to its ability to solubilize a wide range of organic substrates. Acetonitrile, dichloromethane, pyridine or mixtures of these four solvents could also be used. The electrolyte $LiBF_4$ could be used in place of $LiClO_4$ with little decrease in yield, but tetraalkylammonium salts were not competent electrolytes.

Although most of the other mediators studied were inferior to the original NHPI, we reasoned that the addition of electron-withdrawing groups to the phthalamide scaffold would improve the reactivity of the catalyst[24]. Thus, tetrachloro-*N*-hydroxyphthalimide ($Cl_4$NHPI) was chosen, owing to its ease of preparation from tetrachlorophthalic anhydride, an industrial non-toxic flame retardant (which is easily obtained for about US$30 per kilogram from suppliers of chemicals for laboratories). The expectation of increased reactivity was supported by cyclic voltammetry data. In the case of NHPI, a reversible redox couple is observed at 0.78 V versus Ag/AgCl in the presence of excess pyridine, whereas $Cl_4$NHPI shows a redox couple at 0.87 V versus Ag/AgCl under identical conditions. This slightly increased oxidation potential is consistent with the generation of a higher-energy and more-reactive

[1]Department of Chemistry, The Scripps Research Institute, La Jolla, California 92037, USA. [2]Asymchem Life Science (Tianjin), Tianjin Economic-Technological Development Zone, Tianjin 300457, China. [3]Chemical Development, Bristol–Myers Squibb, New Brunswick, New Jersey 08903, USA.
*These authors contributed equally to this work.

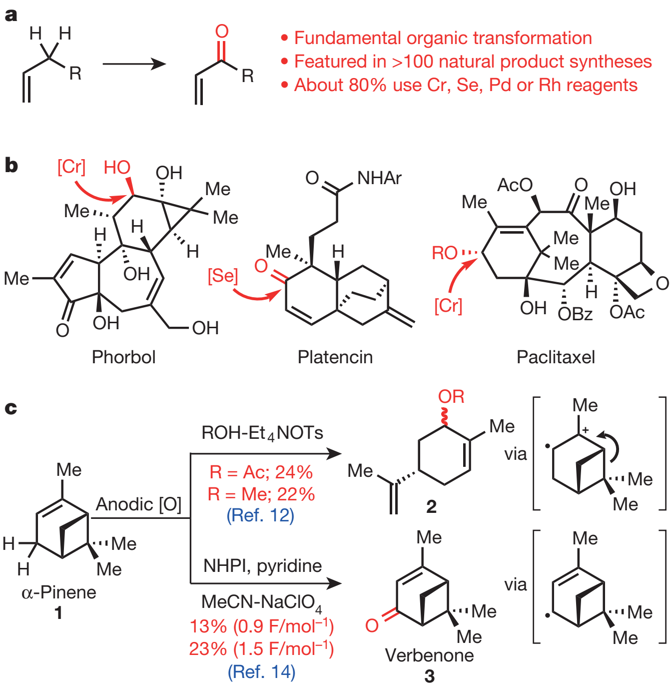



**Figure 1 | Widely applied allylic oxidation. a**, Sustainable allylic C–H oxidation is an unsolved problem. **b**, Case studies from classic total syntheses. **c**, Electrochemical oxidation represents a potential solution.

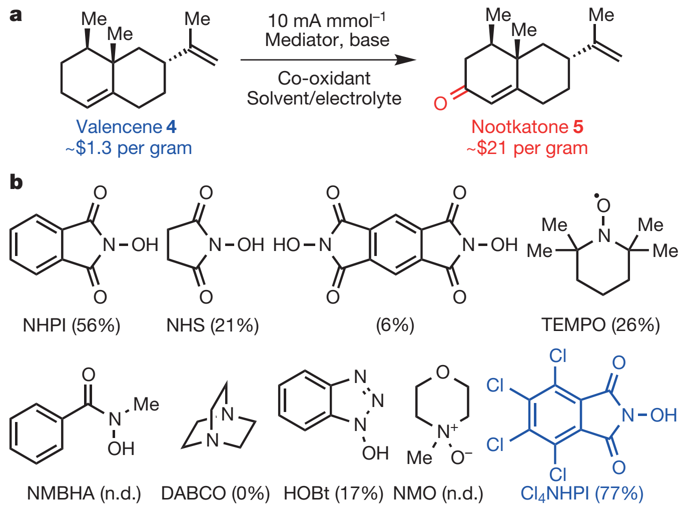



**Figure 2 | Optimization of a sustainable allylic C–H oxidation. a**, Co-oxidants (NHPI, pyridine, $CH_3CN$): air (6%), bubbling $O_2$ (18%), $Bz_2O_2$ (0%), $^tBu_2O_2$ (0%), $H_2O_2$, (27%), $^t$BuOOH (51%), $PhC(CH_3)_2OOH$ (43%). Bases (NHPI, $^t$BuOOH, $CH_3CN$): pyridine (51%), 2,6-lutidine (10%), 2,4,6,-collidine (13%), $Et_3N$ (0%), DBU (0%), $Li_2CO_3$ (trace). Solvents (NHPI, $^t$BuOOH, pyridine): $CH_3CN$ (51%), pyridine (40%), acetone (56%), $CH_2Cl_2$ (21%), MeOH (trace), DMF (5%), DMSO (14%), HFIP (0%), EtOAc (trace), THF (trace). Electrolyte (NHPI, $^t$BuOOH, pyridine, acetone): $LiClO_4$ (56%), $LiBF_4$ (41%), $Et_4NClO_4$ (0%). **b**, Mediators ($^t$BuOOH, pyridine, acetone). Optimized electrochemical parameters: $Cl_4$NHPI (0.2 equiv.), pyridine (2 equiv.), $^t$BuOOH (1.5 equiv.), $LiClO_4$ (0.6 equiv.), acetone (0.16 M in substrate), reticulated vitreous carbon electrodes, 10 mA per mmol of substrate. n.d., not detected.

phthalimido-*N*-oxyl radical, and the use of $Cl_4$NHPI as a mediator led to a cleaner reaction profile and an isolated yield of 77%.

The final optimized conditions for oxidation of **4** to **5** is as follows: 20 mol% $Cl_4$NHPI, pyridine (2.0 equiv.), $^t$BuOOH (1.5 equiv.), and $LiClO_4$ as the supporting electrolyte (0.1 M) in acetone (6 ml per mmol of substrate) under constant-current conditions in an undivided cell. No precautions to exclude oxygen or water were undertaken, and technical-grade solvents and reagents and a simple set-up of two reticulated vitrous carbon electrodes separated by a glass slide were used (see Supplementary Information for a photographic guide of the experimental set-up).

Our initial explorations into the tolerance focused on several cycloalkene-derived substrates relevant to drug discovery (Fig. 3); *tert*-butyl cyclohexenone **6** could be prepared as a single regioisomer in 52% yield, whereas phenyl cyclohexenone **7** was prepared in 55% yield. Cyclopentenone **8** could be prepared from trimethylsilyl (TMS)-protected cyclopentenol in 52% yield. Unprotected tertiary alcohols **9–12** could be prepared from the corresponding aryl-substituted cyclohexenols in reasonable yields. The Lewis-basic pyridine-containing enone **12** was prepared in 60% yield. No alcohol elimination and aromatization was observed, and no allylic transposition occurred; this complimentary reactivity to most chromium-based oxidants and the use of unprotected alcohols is particularly notable. Furthermore, the unsubstituted cyclohexenone products **13** and **14** were prepared in 56% and 51% yield, respectively, comparing favourably to previously reported (see Supplementary Information) $Cr^V$-mediated oxidation for the synthesis of **13**. Substitution at the alpha position of the enone was tolerated as well, with **15** being isolated in 58% yield. In addition to the cyclic substrates described above, several acyclic alkenes of various chain lengths were successfully oxidized under the reaction conditions. Enones **16** and **17** were prepared in 46% and 51% yields, respectively. Propargylic alcohol **18** was formed in 52% yield from its corresponding alkyne.

Owing to their prevalence in the drug discovery, flavour and fragrance industries, our subsequent efforts focused on a variety of representative terpene classes. The oxidation products of monoterpenes are among the most widely used and correspondingly valuable substances not only as fragrances and flavours, but also as building blocks in synthesis. As such, we evaluated the efficiency with which the electrochemical allylic oxidation could be applied to these substances. Verbenone (**3**), previously prepared in 13%–23% yield[14], was prepared in 67% yield. Both isomers of the food additive theaspirane could be oxidized to the natural products *cis*- and *trans*-theaspirone **19** and **20** in 63% and 49% yield, respectively. Carvone-derived enone **21** was prepared in 47% yield, whereas carvone (**22**) was prepared in 42% yield. Myrtenol acetate and nopol acetate were converted to the corresponding oxo-myrtenol and oxo-nopol compounds **23** and **24** in 64% and 43% yield, respectively. Furthermore, aza-nopol analogues **25** and **26** were prepared in 42% and 53% yield, respectively, further highlighting the tolerance for nitrogen-containing functionalities under these reaction conditions. In all cases, isolated yields are comparable to those in the literature using other methods, and low product yields were obtained when using previously reported NHPI/$O_2$ conditions after prolonged heating[16–18]. To put these results in context, an extensive comparative survey of literature conditions and yields is included in Supplementary Information.

Sesquiterpenoid and diterpenoid natural products, many of which are components of essential oils, have provided inspiration for many strategies and methods in synthesis, primarily owing to their complex and dense structures and their promising biological activities[25]. Allylic oxidation of valencene gave nootkatone (**5**) in 77% yield (see above). Sclareolide-derived terpene **27** was synthesized in 75% yield using the electrochemical method. Eudesmane natural products carissone (**28**, as its TMS ether) and cyperone (**29**) were prepared in 54% and 51% yields, respectively. The related aubergenone skeleton was successfully oxidized to give enone **30** in 67% yield. Isolongifolenone (**31**) was prepared from the feedstock chemical isolongifolene in 91% yield. Natural products in the guaiane family were oxidized to afford the natural products pancherione acetate (**32**) and rotundone (**33**) in 41% and 44% yield, respectively, and the complex [3.2.1]-bicyclic system in cedrene was oxidized to give cedren-10-one (**34**) in 53% yield. Under electrochemical conditions, abietic acid derivative **35** could be prepared in 66% isolated yield along with 12% recovered starting material. Although

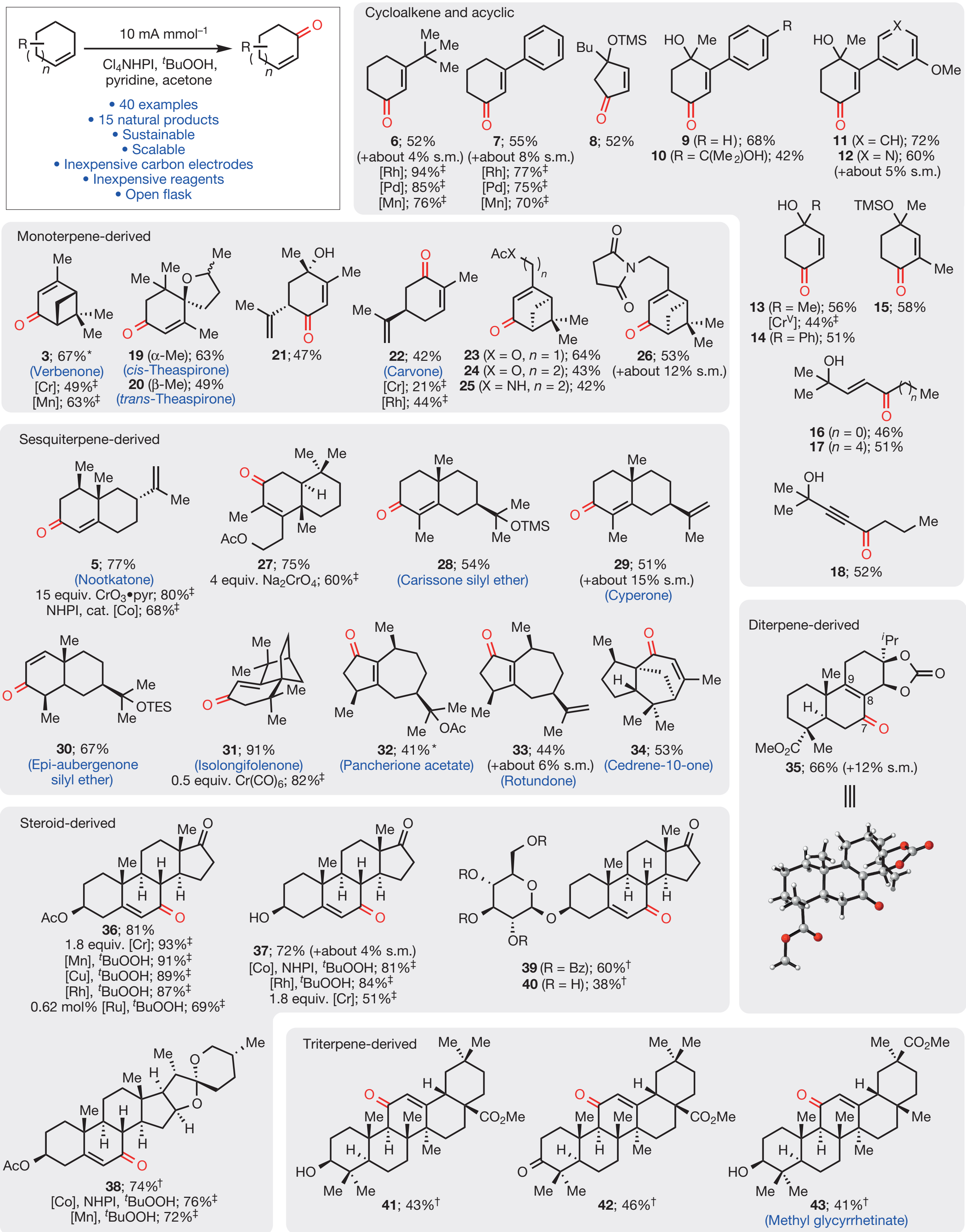



**Figure 3 | Scope of the electrochemical allylic oxidation.** Yields refer to isolated yields of products after chromatography on $SiO_2$. Standard conditions: terpenoid substrate (0.5 mmol), $Cl_4NHPI$ (0.1 mmol), pyridine (1.0 mmol), $^{t}BuOOH$ (0.75 mmol), $LiClO_4$ (0.3 mmol), acetone (3 ml). *MeCN (3 ml) replaced acetone as solvent. †Terpenoid substrate (0.25 mmol), $Cl_4NHPI$ (0.1 mmol), pyridine (1.0 mmol), $LiClO_4$ (0.3 mmol), $CH_2Cl_2$ (1.5 ml) and acetone (1.5 ml) were used. ‡See Supplementary Information for references. s.m., starting material.

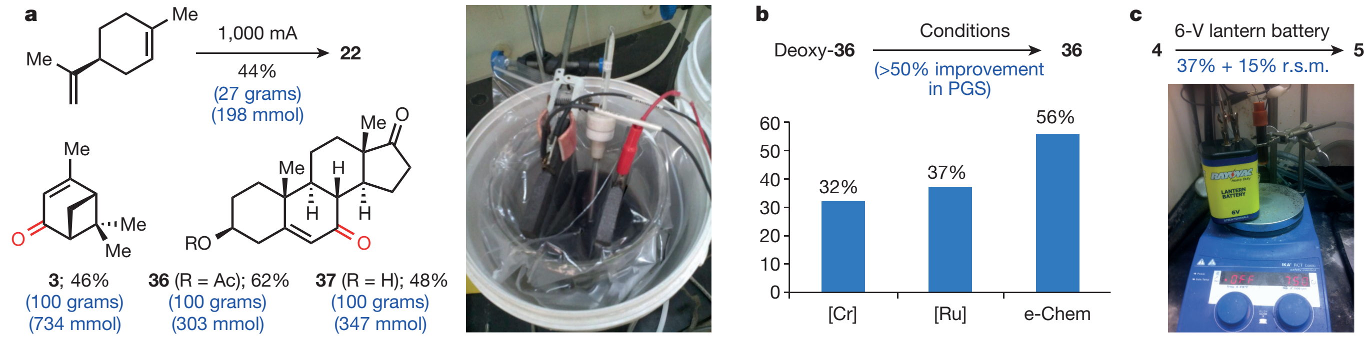



**Figure 4 | Practicality of the electrochemical method. a**, Electrochemical allylic oxidation on a 100-g scale. **b**, Calculated Process Greenness Score (PGS) for $CrO_3$-mediated, $RuCl_3$-catalysed, electrochemical oxidation of deoxy-**36** to **36** shows improvement from 32.1% to 55.8%. Cost and toxicity associated with chromium and ruthenium use and disposal are not included in the PGS. **c**, Use of 6-V lantern battery as a readily available power source for allylic oxidation. r.s.m., recovered starting material.

the starting material contained the $\Delta^{7,8}$ bond, isomerization to the $\Delta^{8,9}$ olefin occurred, and no $\Delta^{7,8}$-enone was observed. This anomalous result may be due to hydrogen-atom abstraction at the C9 position followed by trapping of the allylic radical at the C7 position.

The oxidation of steroid and triterpene substrates has been shown to improve properties such as solubility and pharmacokinetics, and the development of tools to modify their 'oxidation barcodes' are of immediate importance[26]. Electrochemical oxidation of acetate-protected dehydroepiandrosterone (DHEA) gave enone product **36** in 81% yield. Unprotected DHEA could also be oxidized to give enone **37** in 72% yield. Diosgenin acetate underwent smooth oxidation to give **38** in 74% yield. Encouraged by the tolerance for free alcohols and sensitive acetals, glycosylated derivatives of DHEA were evaluated. The tetrabenzoyl protected glycoside **39** was prepared in 60% yield and the unprotected glycoside **40** could be prepared in 38% yield, demonstrating chemoselectivity that would not be possible using classical oxidants such as chromium. Free-hydroxyl-containing methyl oleanolate derivative **41** was prepared in 43% isolated yield, whereas oxidation of the ketone-containing substrate gave the desired enone **42** in 46% yield. Methyl glycyrrhetinate (**43**), an important starting material for a variety of oxidized, medicinally relevant triterpenes, was also prepared in 41% yield.

In nearly all substrates evaluated, isolated yields compare favourably to literature precedent using traditional reagent-based oxidants. In some cases, such as the known conversion of **4** to **5** using 15 equivalents of $CrO_3$•pyridine in 80% yield, not only is the isolated yield comparable, but our conditions obviate the need for the excessive use of toxic reagents and minimize the use of solvent and aqueous media for extraction and isolation.

We demonstrate the feasibility of adopting this technology in a process setting, using the described conditions. These reactions were conducted using inexpensive graphite plate electrodes in a beaker open to air (Fig. 4a, inset photograph), and $LiBF_4$ was used as the supporting electrolyte. Using this set-up, **22** was prepared on a 27-g scale (198 mmol) from limonene in 44% yield, and **3** was prepared in 55% yield on a 27-g scale. Further scale-up to 100 g (734 mmol) gave 46% yield of **3**. Carrying out the identical transformation with traditional chromium reagents (for example, toxic chromium hexacarbonyl) would require at least 81 g of chromium reagent followed by extensive efforts to remove chromium-based contaminants. Sterol **37** and its acetate **36** were produced in 48% yield (100 g, 347 mmol) and 62% yield (100 g, 303 mmol), respectively. Highlights of this successful external field test include operational simplicity, safe procedure, simple workup and ease of product isolation.

To verify the improved environmental footprint of the electrochemical allylic oxidation, we compared the Process Greenness Scores (PGSs) for the electrochemical preparation of **36** to two known literature methods[20,27] (Fig. 4b). The PGS is a method often used by industrial companies to evaluate the potential environmental impact of chemical manufacturing processes. The scoring parameters are closely aligned with the 12 principles of green chemistry[28], among which limiting waste generation and maximizing process efficiency are two main metrics of environmentally friendly processes. A greener reaction has a higher PGS. Oxidation reactions, especially aliphatic C–H oxidations, are generally associated with lower-than-average PGSs due to typically low process yields and the common use of toxic metal mediators in stoichiometric quantities. Unsurprisingly, the $CrO_3$-mediated oxidation of deoxy-**36** scored lowest in terms of PGS (32.1%). The $RuCl_3$-catalysed oxidation had an improved, albeit still modest, PGS of 37.1%. We were pleased to find that the electrochemical allylic oxidation showed a markedly improved PGS of 55.8% (an improvement of >50%). This difference is substantial and shows the step-change in applicability of this new technology. For comparison, the PGS for a typical amide bond formation (EDC, HOBt) ranges from 55% to 70%, whereas the PGS of the widely used palladium-catalysed cross-coupling of aryl halides with boronic acids falls between 45% and 60%. As a further testament to its robustness, the electrochemical oxidation was carried out using a 6-V lantern battery[29,30] (Fig. 4c), with valencene being converted to nootkatone in 37% yield with 15% recovered starting material.

Although this reaction is useful for the oxidation of numerous natural and unnatural carbon skeletons, it is not without its limitations. For example, although cyclic substrates are all reactive, not all acyclic alkenes give very high conversion to enone products, nor do electron-deficient alkenes. In some cases, allylic alcohol products were isolated alongside enone products, although with prolonged reaction times these products are converted to the desired enones. Yields for some substrate classes were modest (for example, **32** and **33**), in part owing to incomplete conversion, substrate decomposition, or adsorption to the electrode surface. However, in nearly all cases isolated yields are comparable to alternative procedures present in the literature.

Mechanistic aspects of the initiation step of this new transformation may have parallels to other NHPI-catalysed oxidations[16–18] (Fig. 5). Thus, deprotonation of $Cl_4NHPI$ by pyridine, followed by anodic

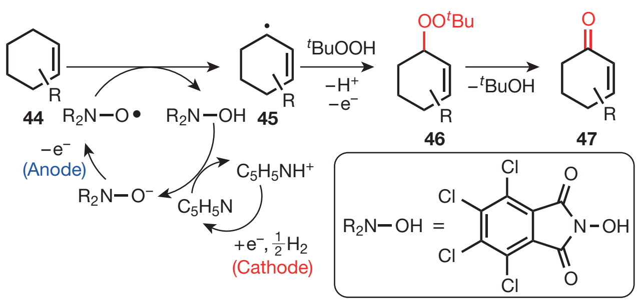



**Figure 5 | Proposed mechanism for electrochemical allylic oxidation.** The boxed structure is the catalyst, which is abbreviated in the catalytic cycle as $R_2N–OH$.

oxidation, leads to the *tetra*-chlorophthalimido *N*-oxyl radical species. Olefinic substrate **44** would then undergo hydrogen atom abstraction, regenerating $Cl_4$NHPI and the relatively stable allylic radical species **45**. Reaction with electrochemically generated $^t$BuOO$^•$ would then give allylic peroxide **46**, which, upon elimination of $^t$BuOH, affords enone **47** (see Supplementary Information for more details).

**Received 7 October 2015; accepted 16 February 2016.**
**Published online 20 April 2016.**

1. Gutekunst, W. R. & Baran, P. S. C–H functionalization logic in total synthesis. *Chem. Soc. Rev.* **40,** 1976–1991 (2011).
2. Weidmann, V. & Maison, W. Allylic oxidations of olefins to enones. *Synthesis* **45,** 2201–2221 (2013).
3. Nakamura, A. & Nakada, M. Allylic oxidations in natural product synthesis. *Synthesis* **45,** 1421–1451 (2013).
4. Sequeira, C. A. C. & Santos, D. M. F. Electrochemical routes for industrial synthesis. *J. Braz. Chem. Soc.* **20,** 387–406 (2009).
5. Degner, D. in *Electrochemistry III* (ed. Steckchan, E.) 1–95 (Springer, 1988).
6. Moeller, K. D. Synthetic applications of anodic electrochemistry. *Tetrahedron* **56,** 9527–9554 (2000).
7. Sperry, J. B. & Wright, D. L. The application of cathodic reductions and anodic oxidations in the synthesis of complex molecules. *Chem. Soc. Rev.* **35,** 605–621 (2006).
8. Yoshida, J.-i., Kataoka, K., Horcajada, R. & Nagaki, A. Modern strategies in electroorganic synthesis. *Chem. Rev.* **108,** 2265–2299 (2008).
9. Francke, R. & Little, R. D. Redox catalysis in organic electrosynthesis: basic principles and recent developments. *Chem. Soc. Rev.* **43,** 2492–2521 (2014).
10. Gütz, C., Bänziger, M., Bucher, C., Galvão, T. R. & Waldvogel, S. R. Development and scale-up of the electrochemical dehalogenation for the synthesis of a key intermediate for NS5A inhibitors. *Org. Process Res. Dev.* **19,** 1428–1433 (2015).
11. Shono, T. & Kosaka, T. Organic synthesis by electrolysis III anodic allylic substitution. *Tetrahedr. Lett.* **9,** 6207–6208 (1968).
12. Shono, T. & Ikeda, A. Electroorganic chemistry X anodic allylic substitution. *J. Am. Chem. Soc.* **94,** 7892–7898 (1972).
13. Masui, M., Hara, S., Ueshima, T., Kawaguchi, T. & Ozaki, S. Anodic oxidation of compounds having benzylic or allylic carbon and $\alpha$-carbon to hetero atom using *N*-hydroxyphthalimide as a mediator. *Chem. Pharm. Bull.* **31,** 4209–4211 (1983).
14. Masui, M., Hosomi, K., Tsuchida, K. & Ozaki, S. Electrochemical oxidations of olefins using *N*-hydroxyphthalimide as a mediator. *Chem. Pharm. Bull.* **33,** 4798–4802 (1985).
15. Ueda, C., Noyama, M., Ohmori, H. & Masui, M. Reactivity of phthalimide-*N*-oxyl: a kinetic study. *Chem. Pharm. Bull.* **35,** 1372–1377 (1987).
16. Foricher, J., Fürbringer, C. & Pfoertner, K. Process for the catalytic oxidation of isoprenoids having allylic groups. US patent 5,030,739 (1991).
17. Ishii, Y. *et al.* A novel catalysis of *N*-hydroxyphthalimide in the oxidation of organic substrates by molecular oxygen. *J. Org. Chem.* **60,** 3934–3935 (1995).
18. Recupero, F. & Punta, C. Free radical functionalization of organic compounds catalyzed by *N*-hydroxyphthalimide. *Chem. Rev.* **107,** 3800–3842 (2007).
19. Miller, R. A., Li, W. & Humphrey, G. R. A ruthenium catalyzed oxidation of steroidal alkenes to enones. *Tetrahedr. Lett.* **37,** 3429–3432 (1996).
20. Harre, M. *et al.* Some reaction safety aspects of ruthenium-catalyzed allylic oxidations of $\Delta$-5-steroids in the pilot plant. *Org. Process Res. Dev.* **2,** 100–104 (1998).
21. Campbell, A. N. & Stahl, S. S. Overcoming the 'oxidant problem': strategies to use $O_2$ as the oxidant in organometallic C–H oxidation reactions catalyzed by Pd (and Cu). *Acc. Chem. Res.* **45,** 851–863 (2012).
22. Osterberg, P. M. *et al.* Experimental limiting oxygen concentrations for nine organic solvents at temperatures and pressures relevant to aerobic oxidations in the pharmaceutical industry. *Org. Process Res. Dev.* **19,** 1537–1543 (2015).
23. Mudryk, B., Zheng, B., Chen, K. & Eastgate, M. D. Development of a robust process for the preparation of high-quality dicyclopropylamine hydrochloride. *Org. Process Res. Dev.* **18,** 520–527 (2014).
24. Cai, Y., Koshino, N., Saha, B. & Espenson, J. H. Kinetics of self-decomposition and hydrogen atom transfer reactions of substitute phthalimide *N*-oxyl radicals in acetic acid. *J. Org. Chem.* **70,** 238–243 (2005).
25. Modzelewska, A., Sur, S., Kumar, S. K. & Khan, S. R. Sesquiterpenes: natural products that decrease cancer growth. *Curr. Med. Chem. Anticancer Agents* **5,** 477–499 (2005).
26. Michaudel, Q. *et al.* Improving physical properties via C–H oxidation: chemical and enzymatic approaches. *Angew. Chem. Int. Ed.* **53,** 12091–12096 (2014).
27. Marwah, P. & Lardy, H. A. Process for effecting allylic oxidation using dicarboxylic acid imides and chromium reagents. US patent 6,384,251 (2002).
28. Anastas, P. T. & Warner, J. C. *Green Chemistry: Theory and Practice* (Oxford Univ. Press, 1998).
29. Frey, D. A., Wu, N. & Moeller, K. D. Anodic electrochemistry and the use of a 6-volt lantern battery: a simple method for attempting electrochemically based synthetic transformations. *Tetrahedr. Lett.* **37,** 8317–8320 (1996).
30. Frankowski, K. J., Liu, R., Milligan, G. L., Moeller, K. D. & Aubé, J. Practical electrochemical anodic oxidation of polycyclic lactams for late stage functionalization. *Angew. Chem. Int. Ed.* **54,** 10555–10558 (2015).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** This work was supported by an NSF predoctoral fellowship (B.R.R.), National Institute of General Medical Sciences grant GM-097444, Asymchem and Bristol–Myers Squibb. We thank D.-H. Huang and L. Pasternack for assistance with NMR spectroscopy; A. L. Rheingold, C. E. Moore and M. A. Galella for X-ray crystallographic analysis; and D. G. Blackmond, O. Luca, T. Paschkewitz, Y. Ishihara and T. Razler for discussions.

**Author Contributions** E.J.H., B.R.R. and P.S.B. conceived this work; E.J.H., B.R.R., K.C., M.D.E. and P.S.B. designed the experiments; E.J.H. and B.R.R. conducted the experiments and analysed the data; Y.C. and J.T. performed the large-scale experiments; E.J.H., B.R.R., K.C., M.D.E. and P.S.B. wrote the manuscript.

**Author Information** Metrical parameters for the structure of **24** are available free of charge from the Cambridge Crystallographic Data Centre under reference number CCDC-1058554. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to P.S.B. (pbaran@scripps.edu).

# LETTER

doi:10.1038/nature17434

# Chondritic xenon in the Earth's mantle

Antonio Caracausi[1,2], Guillaume Avice[2], Peter G. Burnard[2]‡, Evelyn Füri[2] & Bernard Marty[2]

**Noble gas isotopes are powerful tracers of the origins of planetary volatiles, and the accretion and evolution of the Earth. The compositions of magmatic gases provide insights into the evolution of the Earth's mantle and atmosphere[1–7]. Despite recent analytical progress in the study of planetary materials[8,9] and mantle-derived gases[2–7], the possible dual origin[1,10] of the planetary gases in the mantle and the atmosphere remains unconstrained. Evidence relating to the relationship between the volatiles within our planet and the potential cosmochemical end-members is scarce[5]. Here we show, using high-precision analysis of magmatic gas from the Eifel volcanic area (in Germany), that the light xenon isotopes identify a chondritic primordial component that differs from the precursor of atmospheric xenon. This is consistent with an asteroidal origin for the volatiles in the Earth's mantle, and indicates that the volatiles in the atmosphere and mantle originated from distinct cosmochemical sources. Furthermore, our data are consistent with the origin of Eifel magmatism being a deep mantle plume. The corresponding mantle source has been isolated from the convective mantle since about 4.45 billion years ago, in agreement with models that predict the early isolation of mantle domains[11]. Xenon isotope systematics support a clear distinction between mid-ocean-ridge and continental or oceanic plume sources[6], with chemical heterogeneities dating back to the Earth's accretion[1,7]. The deep reservoir now sampled by the Eifel gas had a lower volatile/refractory (iodine/plutonium) composition than the shallower mantle sampled by mid-ocean-ridge volcanism, highlighting the increasing contribution of volatile-rich material during the first tens of millions of years of terrestrial accretion.**

Owing to their inertness, low abundances and the presence of several different radioactive chronometers in their isotope systematics, the noble gases are excellent geochemical tracers of the formation and subsequent evolution of the Earth[1–7]. However, the origin of terrestrial noble gases is not fully understood. The isotopic composition of atmospheric xenon (Xe) is particularly puzzling because it appears to be strongly isotopically fractionated with respect to solar (derived from the protosolar nebula gas and represented by the solar wind composition) and chondritic (derived from an asteroid-like reservoir) components (see, for example, ref. 12). This feature could be a result of ancient atmospheric escape processes, but even after correction for mass-dependent isotope fractionation the isotope composition of atmospheric Xe cannot easily be related to a chondritic or solar origin[12]. One way to investigate the origin of terrestrial volatiles is to precisely document the compositions of noble gases that have been stored in the terrestrial mantle, presumably since the formation of the Earth.

Mantle-derived $CO_2$-rich gases are particularly powerful resources for investigating mantle-derived noble gases because the large quantities of sample material available make high-precision measurements possible[2–5,13]. Here we report Xe isotopic measurements in gases from a $CO_2$-rich well (Victoriaquelle) in the Eifel volcanic region (Germany). Geophysical and geochemical evidence suggests that the Eifel volcanism, which took place from 700 kyr ago to 11 kyr ago, was related to continental rifting and large-scale mantle upwelling[14–17]. The Victoriaquelle well, in the southwest of the Eifel region, emits $CO_2$-dominated gases (99.7%–99.8% $CO_2$) with helium isotope ratios of 4.2–4.5 Ra (where Ra is the helium isotope signature of the Earth's atmosphere) and $^{40}Ar/^{36}Ar$ ratios of up to 2,690 (ref. 18), consistent with low levels of atmospheric contamination and predominantly mantle-derived volatile emissions[18].

Our Xe isotope data (normalized to $^{130}Xe$; Fig. 1, Extended Data Table 1) demonstrate that there is a mantle-derived component to this noble gas, marked by a 2.45% excess (relative to air) of $^{129}Xe$ from the decay of extinct $^{129}I$ (half-life of 16 Myr). The dataset also highlights an excess of the lightest isotopes ($^{124–128}Xe$) relative to air Xe. Because the light Xe isotopes are not affected by radiogenic or fissiogenic production, this excess $^{124–128}Xe$ must represent a primordial Xe component. Excesses of light Xe isotopes have already been recognized in some mantle-derived gases[4,5,13]. Furthermore, given that isotopic fractionation during mantle processing is unlikely, these light Xe spectra must therefore reflect the presence of either solar or chondritic Xe (either average

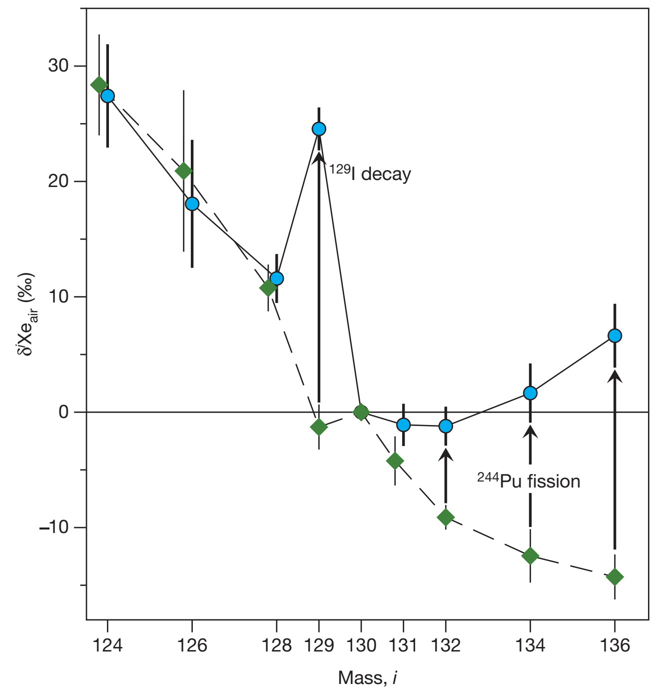



**Figure 1 | Xe isotope composition of the Victoriaquelle gas.** Data (blue filled circles) are normalized to the isotope composition of atmospheric Xe and to $^{130}Xe$. Deviations from the atmospheric composition ($Xe_{air}$) are expressed in delta notation as parts per mil (‰). For comparison, we show the composition of a mixture composed of 84% atmospheric Xe and 16% chondritic (average carbonaceous chondrite) Xe (green diamonds; see Methods for the derivation of the component fractions). The excesses at masses $i = 129$ and $i = 131–136$ are the products of the extinct radioactivity of $^{129}I$ (half-life of 16 Myr) and $^{244}Pu$ (half-life of 82 Myr), respectively. Error bars indicate $\pm 1\sigma$.

[1]Instituto Nazionale di Geofisica e Vulcanologia, Sezione di Palermo, 90146 Palermo, Italy. [2]Centre de Recherches Pétrographiques et Géochimiques, UMR 7358, Université de Lorraine, CNRS, 54501 Vandoeuvre-lès-Nancy, France.
‡Deceased.

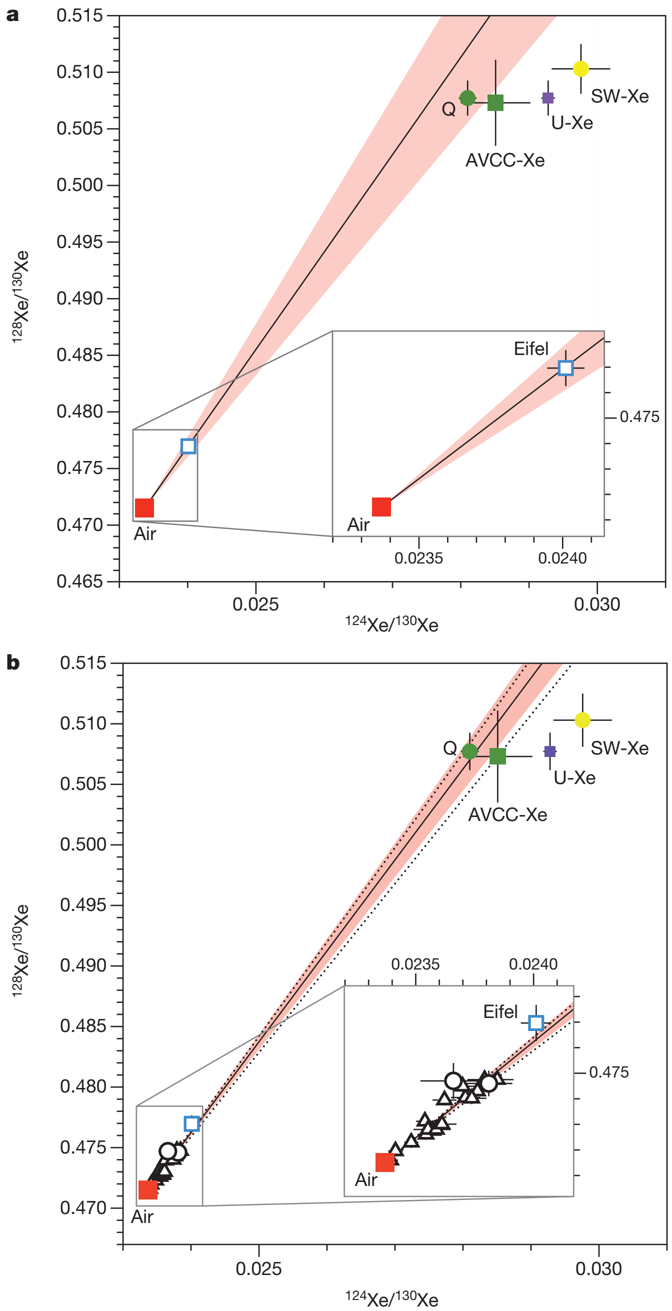



**Figure 2 | Light Xe isotope correlations. a**, The Eifel composition (open blue square) was derived from the mean of 15 measurements on three different aliquots of the same gas (see Methods). The solid black line is a best-fit line through the air (filled red square) and Eifel compositions; correlation errors were computed using the Isoplot code (courtesy of K. Ludwig, Berkeley Geochronology Center) (the light red envelope represents $\pm 1\sigma$ error). The compositions of phase-Q Xe ('Q', filled green circle; the major carrier of heavy noble gases trapped in primitive meteorites), average-carbonaceous-chondrite Xe ('AVCC-Xe', filled green square; ref. 9 and references therein), solar-wind Xe ('SW-Xe', filled yellow square; ref. 8) and the inferred progenitor of atmospheric Xe ('U-Xe', filled purple square; ref. 9) are also shown for comparison (with $\pm 1\sigma$ error bars). The correlation extrapolates to a chondritic, rather than solar or U-Xe, end-member in the mantle. Because of the overlap in the compositions of the 'Q'-type gases and average carbonaceous chondrites, it is not possible to distinguish the nature of the chondritic Xe carrier phase in the accreting Earth (in our discussion we use 'AVCC' without distinguishing 'Q' from the bulk chondritic composition). **b**, Comparison with other $CO_2$-rich well gases (Bravo Dome, black triangles; Harding County and Caroline, open circles; refs 5, 13, 19). The Eifel gas is seen either to contain the highest proportion of primordial Xe or to have been less affected by air contamination. The best-fit correlation for the Bravo Dome dataset (black dashed lines represent the upper and lower limit with $\pm 1\sigma$ error range) points to a chondritic composition for Xe in the upper mantle. The best-fit line obtained for the whole dataset (this study and the published well-gas data) also points to a chondritic Xe composition, demonstrating the ubiquitous presence of this component in the mantle.

carbonaceous chondrite or 'phase-Q', a ubiquitous noble gas component found in all classes of primitive meteorites; see, for example, ref. 9) in the Earth's mantle. Because the differences between the isotopic patterns of chondritic and solar-wind Xe are subtle[8,9], it has not previously been possible to differentiate between these two potential primordial sources of Xe. Given the nature of our sample, the high analytical precision of the present study and the recently improved precision on the isotope composition of solar-wind Xe (ref. 8), we are able to assign a chondritic rather than solar origin to the Eifel primordial Xe end-member (Fig. 2a, Methods, Extended Data Fig. 1). According to our calculations for the light isotopes, this chondritic Xe component represents 16% ± 2% of the Eifel gas, the remainder being atmospheric in origin (Methods, Extended Data Fig. 2). This is the highest proportion of primordial Xe identified in mantle-derived volatiles so far (Fig. 2b). Furthermore, the fact that all $CO_2$ well gases[4,5,13] analysed so far (from Bravo Dome (USA), Harding County (USA) and Caroline (Australia), and which have been ascribed an upper-mantle origin) also point to a chondritic Xe component and lie on a single correlation line with the Eifel gas (shown below to have a mantle plume origin; Fig. 2b), demonstrates the existence of a ubiquitous primordial Xe component in the Earth's mantle. Therefore, chondritic Xe was widely distributed in the proto-mantle during the Earth's accretion. Krypton isotopes also point to a chondritic source for Bravo Dome gases sampling the upper mantle[5], and, together with the present Xe data, support an asteroidal origin for heavy noble gases in the whole mantle.

This study points to several sources of volatile elements on Earth. The ancestor of atmospheric Xe was neither chondritic nor solar in origin because it had to have been relatively depleted in the heavy Xe isotopes (notably $^{134}$Xe and $^{136}$Xe) compared to documented primordial Xe components[19]. Known nuclear processes cannot resolve this issue because they can only contribute, not deplete, these isotopes. This problem, first recognized four decades ago[20], led to the definition of a primitive Xe component dubbed 'U-Xe' (not to be confused with Xe isotopes produced by $^{238}$U fission), which was of solar composition for the light isotopes and depleted in both $^{134}$Xe and $^{136}$Xe relative to solar and chondritic Xe (ref. 19). Thus, two Xe components appear to co-exist on Earth: chondritic Xe preserved in the mantle and U-Xe found in the atmosphere. Consequently, the non-radiogenic, non-fissiogenic Xe in the atmosphere cannot have been derived from the mantle. To prevent mixing between the two components, the atmospheric Xe must have been added after growth of the Earth had largely been completed.

The heavy Xe isotope composition ($^{131,132,134,136}$Xe) of the mantle is more complex, being a mixture of four isotopically distinct end-members: (1) atmospheric Xe; (2) primordial Xe; (3) fissiogenic Xe produced from $^{244}$Pu ($^{\mathrm{Pu}}$Xe); and (4) fissiogenic Xe derived from $^{238}$U ($^{\mathrm{U}}$Xe). $^{244}$Pu and $^{238}$U each produce fissiogenic Xe isotopes in characteristic proportions, which differ from those of atmospheric or chondritic Xe. Excesses of fissiogenic or radiogenic $^{131-136}$Xe and $^{129}$Xe in natural samples can be used to distinguish between magmatic sources and to constrain the timing of mantle differentiation. Both $^{129}$Xe and $^{136}$Xe were produced in the early Earth by decay of extinct radiochronometers—$^{129}$I (half-life of 16 Myr) decaying to $^{129}$Xe and $^{244}$Pu (half-life of 80 Myr) producing $^{131-136}$Xe—while extant $^{238}$U also produced $^{131-136}$Xe, but with different ratios to those produced by $^{244}$Pu. Thus, the U-Xe system evolved over the entire history of the Earth, whereas the I-Xe and Pu-Xe systems reflect elemental fractionation that occurred during

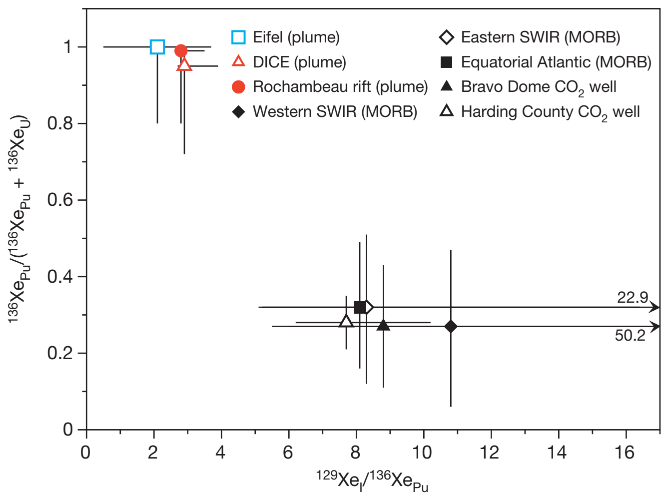



**Figure 3 | Differences in the Xe isotopic compositions of the MORB and mantle plume reservoirs.** The error-weighted ratios of iodine-derived Xe and plutonium-derived Xe ($^{129}Xe_I/^{136}Xe_{Pu}$) versus the fraction of fissiogenic Xe derived from plutonium ($^{136}Xe_{Pu}/(^{136}Xe_{Pu}+^{136}Xe_U)$) enable plume-type mantle sources to be resolved from MORB-type mantle sources[6,7] ($\pm 1\sigma$ error bars). The $^{129}Xe_I/^{136}Xe_{Pu}$ and $^{136}Xe_{Pu}/(^{136}Xe_{Pu}+^{136}Xe_U)$ ratios of worldwide MORB, plume sources and $CO_2$ well gas (all data except Eifel) are from refs 6 and 7, computed assuming an average-carbonaceous-chondrite primordial component for all data. The data for the convective mantle (equatorial Atlantic MORB, filled black square; western Southwest Indian Ridge (SWIR), filled black diamond; eastern SWIR, open black diamond; Harding County gas, open black triangle; Bravo Dome gas, filled black triangle) and the mantle plume sources (Iceland (DICE), open red triangle; Rochambeau rift, filled red circle; Eifel gas, open blue square) define two distinct fields in this diagram, highlighting the different histories and compositions of their respective reservoirs.

only the first 100 Myr and 500 Myr, respectively. The fissiogenic Xe isotope composition, obtained after correction for the atmospheric Xe contribution and by assuming a chondritic Xe composition for primordial Xe (Methods, Extended Data Figs 3, 4), suggests that excesses of heavy Xe isotopes resulted from $^{244}Pu$ fission rather than $^{238}U$ fission (Methods, Extended Data Figs 5, 6). Quantitatively, the fissiogenic Xe contribution to the Eifel gases is 2.26% ± 0.28%, with the remainder being atmospheric or primordial (Methods). Previous estimates of the proportion of $^{238}U$- versus $^{244}Pu$-derived Xe in the mantle depended on the initial Xe isotope composition of the mantle[8] (chondritic or solar); the fact that we demonstrate that the light Xe isotopes are chondritic in origin (Fig. 2a, Extended Data Fig. 1) allows us to confidently establish a $^{Pu}Xe/(^{Pu}Xe+^{U}Xe)$ ratio of 0.8–1.0 ($\pm 1\sigma$) for the Eifel mantle source (Fig. 3, Methods).

In comparison, the other $CO_2$-rich well gases (Bravo Dome and Harding County; refs 4, 5, 13) have significantly lower $^{Pu}Xe/(^{Pu}Xe+^{U}Xe)$ ratios of 0.06–0.51 ($\pm 1\sigma$) (Fig. 3). Their mantle source has been identified as the convective upper mantle, which also supplies magmas to mid-ocean ridges worldwide. Mid-ocean ridge basalts (MORBs) that have been analysed with sufficient precision also display low $^{Pu}Xe/(^{Pu}Xe+^{U}Xe)$ ratios, comparable to the $CO_2$-rich well gases above, and define a well-homogenized convective mantle composition that is depleted in $^{Pu}Xe$ isotopes relative to $^{U}Xe$ (ref. 6). In a closed-system mantle with a chondritic Pu/U ratio[21], $^{Pu}Xe$ should dominate over $^{U}Xe$ ($^{Pu}Xe/(^{Pu}Xe+^{U}Xe)=0.97$). A mantle source degassed over geological timescales would see progressive depletion of $^{Pu}Xe$ and concurrent enrichment in $^{U}Xe$ still being produced. Therefore, the Eifel $^{Pu}Xe/(^{Pu}Xe+^{U}Xe)$ ratio—which is close to 1 and higher than the convective mantle ratio of 0.3 (Fig. 3)—suggests that the Eifel mantle source was much less degassed than the MORB mantle source and, hence, less affected by mantle convection through time. This observation is consistent with the source of the Eifel gas being a deep mantle plume. Two other samples associated with mantle plumes (Iceland plume and the Rochambeau rift in the western Pacific[6]) display comparable $^{Pu}Xe/(^{Pu}Xe+^{U}Xe)$ ratios (Fig. 3), also pointing to a mantle plume origin for the Eifel volcanism.

We also calculated a $^{129}Xe_I/^{136}Xe_{Pu}$ ratio of $2.1 \pm 1.6$ ($\pm 1\sigma$) for the Eifel gas (Fig. 3, Methods), a value comparable to the other plume-like signatures (that is, Iceland and the Rochambeau rift[6]) and different from values characteristic of the convective mantle (>5.1; Fig. 3). Assuming a bulk silicate Earth iodine content of between 3 parts per billion (p.p.b.) and 13 p.p.b. (ref. 22 and references therein), we calculated a I-Pu-Xe 'closure age'[12] for the Eifel mantle of 82–139 Myr after the start of Solar System formation, that is, about 4.45 Gyr ago (Methods, Extended Data Fig. 7). Closure ages should be considered as discrete approximations of a continuous process: they assume that the reservoir was open to Xe loss before that time and that it quantitatively retained Xe isotopes produced by extinct radioactivities afterwards. The early closure age calculated here indicates that degassing of the mantle plume source must have been very efficient when the $^{129}I$ and $^{244}Pu$ extinct isotopes were still alive, that is, during the first 100 Myr of the history of the Earth. After this time, the Eifel mantle source became efficiently isolated from mantle convection, thus preserving a high $^{Pu}Xe/(^{Pu}Xe+^{U}Xe)$ ratio comparable within uncertainty to the closed-system value of 0.97. In contrast, the MORB source reservoir continued to lose $^{Pu}Xe$ while at the same time producing $^{U}Xe$ from long-lived $^{238}U$ fission, resulting in a $^{Pu}Xe/(^{Pu}Xe+^{U}Xe)$ ratio of only 0.3 (Fig. 3). Interestingly, a closure age range of 82–139 Myr is consistent with differentiation times of <150 Myr after the start of Solar System formation recorded by the $^{146}Sm$–$^{142}Nd$ (see, for example, ref. 11) and $^{182}Hf$–$^{182}W$ (see, for example, ref. 23) extinct radioactivity systems. Therefore, the Eifel closure ages might date the last large-scale melting events of the proto-Earth.

If the I/Pu ratio was homogenous during the Earth's accretion, then the higher $^{129}Xe/^{136}Xe_{Pu}$ ratio of the MORB-type sources (Fig. 3) would imply an earlier closure age for the upper mantle than for the plume-type mantle (Extended Data Fig. 7). However, noble gas isotope systematics indicate that the plume-type source is less degassed than the MORB reservoir. It would be paradoxical to suggest that the more-degassed MORB-type mantle became closed to the loss of volatiles before the less-degassed mantle plume source[7]. Thus, it seems more likely that the I/Pu ratio was heterogeneous[1,7] during accretion, with a higher I/Pu ratio in the MORB reservoir than in the plume source. The initial I/Pu ratio must have been at least 3.5 times higher in the MORB source (Extended Data Fig. 7) for the upper mantle to have a younger closure age than the lower mantle. Because iodine is a volatile element and plutonium is a refractory element, the increase in the I/Pu ratio from the deep mantle reservoir source of the Eifel gas to the shallow convective mantle can be viewed as a progressive contribution of volatile-rich material to an initially dry proto-Earth.

The results of this study, coupled with published data[4–7,13], indicate that Xe isotopes in the Eifel gas have preserved a chemical signature that is characteristic of other mantle plume sources (Fig. 3). This corroborates the presence of a deep mantle plume source for the Eifel volcanism, as has previously been suspected[16,17]. Although the helium isotopic signature of the Eifel gas (<6 Ra; refs 17, 18) lies within the field of 'low $^{3}He/^{4}He$' mantle plumes (see, for example, ref. 24), the neon isotopes and neon–argon isotope systematics of the volcanic products point to a deep mantle source below the Eifel region[16,17]. The presence of a mantle plume is also supported by geophysical data[15,16,25,26]. Notably, tomographic images show a low-velocity structure at depths of 660–2,000 km, representing deep mantle upwelling under central Europe, that may feed smaller upper-mantle plumes (such as Eifel, Germany and Massif Central, France)[16].

Our results have implications for both the origin of terrestrial volatiles and the mechanisms and timing of their delivery. Neon, and presumably helium, has a solar-like origin[27], suggesting that these gases

were trapped early during terrestrial accretion, before dissipation of the nebular gas. The other mantle noble gases, krypton[5], Xe (this work) and presumably argon, were delivered together with major volatiles such as hydrogen and nitrogen (ref. 28) by asteroidal material before mantle 'closure' (<60 Myr after the start of Solar System formation). Although the non-radiogenic, non-fissiogenic isotope composition of Xe appears to be homogeneous between the deep mantle and the shallower convective mantle, the volatile/refractory (I/Pu) ratio increased during the Earth's accretion, as is independently suggested by palladium–silver isotope data[29]. This is consistent with the existence of a thermal gradient in the forming Solar System, with the innermost zones being too hot to allow condensation of volatile elements during the initial stages of the Earth's accretion. Dissipation of heat over time and/or contributions of volatile-rich bodies from larger heliocentric distances enabled more-efficient trapping of volatile elements in the shallower regions of the growing Earth[30]. The deepest regions of the mantle, now sampled by mantle plumes, have remained efficiently isolated from mantle convection since about 4.45 Gyr ago, thereby preserving a record of the early stages of terrestrial accretion. The origin of the progenitor of atmospheric Xe (U-Xe) remains enigmatic. It is possible that it was added only after the Earth's completion (>82–139 Myr after the start of Solar System formation), thus avoiding any mixing with mantle (chondritic) Xe. This exotic component may have been carried by volatile-rich bodies from the outer Solar System during late veneer episodes or the Late Heavy Bombardment.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 12 August 2015; accepted 8 February 2016.**
**Published online 25 April 2016.**

1. Marty, B. Neon and xenon isotopes in MORB: implications for the Earth–atmosphere evolution. *Earth Planet. Sci. Lett.* **94,** 45–56 (1989).
2. Ballentine, C. J., Schoell, M., Coleman, D. & Cain, B. A. 300-Myr-old magmatic $CO_2$ in natural gas reservoirs of the west Texas Permian basin. *Nature* **409,** 327–331 (2001).
3. Ballentine, C. J., Marty, B., Sherwood Lollar, B. & Cassidy, M. Neon isotopes constrain convection and volatile origin in the Earth's mantle. *Nature* **433,** 33–38 (2005).
4. Holland, G. & Ballentine, C. J. Seawater subduction controls the heavy noble gases composition of the mantle. *Nature* **441,** 186–191 (2006).
5. Holland, G., Cassidy, M. & Ballentine, C. J. Meteorite Kr in the Earth's mantle suggests a late accretionary source for the atmosphere. *Science* **326,** 1522–1525 (2009).
6. Parai, R. & Mukhopadhyay, S. The evolution of MORB and plume mantle volatile budgets: constraints from fission Xe isotopes in Southwest Indian Ridge basalts. *Geochem. Geophys. Geosyst.* **16,** 719–735 (2015).
7. Mukhopadhyay, S. Early differentiation and volatile accretion recorded in deep-mantle neon and xenon. *Nature* **486,** 101–104 (2012).
8. Meshik, A., Hohenberg, C., Pravdivtsev, O. & Burnett, D. Heavy noble gases in solar wind delivered by Genesis mission. *Geochim. Cosmochim. Acta* **127,** 326–347 (2014).
9. Ott, U. Planetary and pre-solar noble gases in meteorites. *Chem. Erde* **74,** 519–544 (2014).
10. Harper, C. L. Jr & Jacobsen, S. B. Noble gases and Earth's accretion. *Science* **273,** 1814–1818 (1996).
11. Boyet, M. & Carlson, R. W. $^{142}$Nd evidence for early (>4.53 Ga) global differentiation of the silicate Earth. *Science* **309,** 576–581 (2005).
12. Porcelli, D. & Ballentine, C. J. Models for distribution of terrestrial noble gases and evolution of the atmosphere. *Rev. Mineral. Geochem.* **47,** 411–480 (2002).
13. Caffee, M. W. *et al.* Primordial noble gases from Earth's mantle: identification of a primitive volatile component. *Science* **285,** 2115–2118 (1999).
14. Hoernle, K., Zhang, Y. S. & Graham, D. Seismic and geochemical evidence for large-scale mantle upwelling beneath the eastern Atlantic and western and central Europe. *Nature* **374,** 34–39 (1995).
15. Wedepohl, K. H. & Baumann, A. Central European Cenozoic plume volcanism with OIB characteristics and indications of a lower mantle source. *Contrib. Mineral. Petrol.* **136,** 225–239 (1999).
16. Goes, S., Spakman, W. & Bijwaard, H. A lower mantle source for central European volcanism. *Science* **286,** 1928–1931 (1999).
17. Buikin, A. *et al.* Noble gas isotopes suggest deep mantle plume source of late Cenozoic mafic alkaline volcanism in Europe. *Earth Planet. Sci. Lett.* **230,** 143–162 (2005).
18. Bräuer, K., Kämpf, H., Niedermann, S. & Strauch, G. Indications for the existence of different magmatic reservoirs beneath the Eifel area (Germany): a multi-isotope (C, N, He, Ne, Ar) approach. *Chem. Geol.* **356,** 193–208 (2013).
19. Pepin, R. O. & Porcelli, D. Xenon isotope systematic, giant impacts, and mantle degassing on the Earth. *Earth Planet. Sci. Lett.* **250,** 470–485 (2006).
20. Takaoka, N. An interpretation of general anomalies of xenon and the isotopic composition of primitive xenon. *J. Mass Spectrosc. Soc. Japan* **20,** 287–302 (1972).
21. Hudson, G. B., Kennedy, B. M., Podosek, F. A. & Hohenberg, C. M. The early Solar System abundance of $^{244}$Pu as inferred from the St. Severin chondrite. *Proc. Lunar Planet. Sci. Conf.* **19,** 547–557 (1989).
22. Avice, G. & Marty, B. The iodine–plutonium–xenon age of the Moon–Earth system revisited. *Phil. Trans. R. Soc. A* **372,** 20130260 (2014).
23. Touboul, M., Puchtel, I. S. & Walker, R. J. $^{128}$W evidence for long-term preservation of early mantle differentiation products. *Science* **335,** 1065–1069 (2012).
24. Graham, D. W. Noble gas isotope geochemistry of mid-ocean ridge and ocean island basalts: characterization of mantle source reservoirs. *Rev. Mineral. Geochem.* **47,** 247–317 (2002).
25. Budweg, M., Bock, G. & Weber, M. The Eifel plume—imaged with converted seismic waves. *Geophys. J. Int.* **166,** 579–589 (2006).
26. Ritter, J. R. R. in *Mantle Plumes: A Multidisciplinary Approach* (eds Ritter, J. R. R. & Christensen, U. R.) 379–404 (Springer-Verlag, 2007).
27. Honda, M. *et al.* Possible solar noble-gas component in Hawaiian basalts. *Nature* **349,** 149–151 (1991).
28. Marty, B. The origins and concentrations of water, carbon, nitrogen and noble gases on Earth. *Earth Planet. Sci. Lett.* **313–314,** 56–66 (2012).
29. Schönbächler, M., Carlson, R. W., Horan, M. F., Mock, T. D. & Hauri, E. H. Heterogeneous accretion and the moderately volatile element budget of Earth. *Science* **328,** 884–887 (2010).
30. Morbidelli, A. *et al.* Source regions and timescales for the delivery of water to the Earth. *Meteorit. Planet. Sci.* **35,** 1309–1320 (2000).

**Acknowledgements** This work is dedicated to Peter G. Burnard, who passed away after the submission of the manuscript. This study was supported by the Instituto Nazionale di Geofisica e Vulcanologia, by the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013 grant agreement no. 267255) and by the Deep Carbon Observatory. D. L. Hamilton helped in setting up the new mass spectrometry system at CRPG. This is CRPG contribution #2413.

**Author Contributions** A.C., P.G.B. and B.M. designed the study. A.C. collected the samples, performed the experiments and analysed the data. G.A. processed the data and wrote the section on the processing procedure in Methods. A.C., P.G.B., G.A. and B.M. wrote the paper. E.F. collected the samples. All authors contributed to the interpretation and discussion of the data and provided comments on and input to the manuscript.

**Author Information** Data obtained in this study are available at the EarthChem library (http://dx.doi.org/10.1594/IEDA/100582). Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to A.C. (antonio.caracausi@ingv.it).

## METHODS

**Analytical method.** High-precision Xe isotopic ratios were determined in the Noble Gas Laboratory at the Centre de Recherches Pétrographiques et Géochimiques (Nancy, France) using multi-collection mass spectrometry. Xe isotopic compositions were determined in a sample of free gas collected from the Victoriaquelle well in the Eifel volcanic district (Germany). Gas samples were collected in pre-evacuated ($10^{-5}$ Pa) steel bottles equipped with a high-vacuum valve at the end, after thorough flushing of connecting tubes with the well gas. We purified and analysed three aliquots of gas (Extended Data Table 1). Active gases were removed by sequential exposure to hot and cold SAES getters. Xe was condensed on a cold finger at liquid-nitrogen temperature and the abundances of all Xe isotopes were measured on a Thermofisher noble-gas multi-collector mass spectrometer (Helix MC Plus) operating in a combination of multi-collection and peak-jumping modes.

We carried out a total of 15 measurements on the three aliquots of the same gas (Extended Data Table 1; errors are $1\sigma/\sqrt{n}$, where $n$ is the number of duplicate measurements). Procedural blanks were performed before and after each measurement. Xe blanks were typically 0.16% of the $^{130}$Xe signal. Therefore, blank corrections were unnecessary and were not applied to the abundances or isotope ratios reported in Extended Data Table 1. Xenon standard runs were analysed before and during the Victoriaquelle analytical session (total of 30 standard runs with $5.37\times10^{-15}$ mol of $^{132}$Xe per aliquot).

We also purified a different aliquot of gas to measure the Ar isotopic ratio. Ar isotope compositions were measured on a GV-instruments multi-collector mass spectrometer. We determined a $^{40}$Ar/$^{36}$Ar ratio of 1,780 for our Victoriaquelle sample, overlapping values reported in previous investigations[18].

**Residuals of the fit on light isotopes.** We performed a series of calculations to quantitatively identify the best candidate (Q-Xe, AVCC-Xe or SW-Xe) for the primordial Xe component measured in the Eifel gas. We first calculated the relative percentages of atmosphere and primordial component required to obtain the measured isotopic ratios ($^{i}$Xe/$^{130}$Xe, $i=124$–128) for each potential primordial component. The calculated proportions were typically around 87% air mixed with 13% primordial Xe. By taking the mean percentages for each light isotope, we then determined the corresponding isotopic compositions of different mixtures of atmosphere and each primordial component. Extended Data Figure 1 depicts the residuals of this mixing for each case (Q-Xe, AVCC-Xe or SW-Xe). These residuals correspond to the differences between the isotopic ratios measured in the Eifel gas and the modelled isotopic ratios, divided by the corresponding Eifel isotopic ratios for normalization and representation purposes. SW-Xe is a poor candidate for the primordial component, largely owing to its high residual for $^{124}$Xe. Q-Xe or AVCC-Xe are the best candidates.

**Deconvolution of the Xe isotope spectrum.** We assumed that the isotopic spectrum of Xe (excluding $^{129}$Xe) was produced from a mixture of four end-members: (1) modern atmosphere; (2) a primordial component (in the following calculations we used Q-Xe as a proxy for present-day bulk chondrite because present-day bulk chondrite analyses (AVCC) probably contain fissiogenic Xe contributions); (3) fissiogenic Xe derived from $^{244}$Pu ($^{\mathrm{Pu}}$Xe); and (4) fissiogenic Xe derived from $^{238}$U ($^{\mathrm{U}}$Xe).

To estimate the contribution of each component, we divided the problem into two stages.

First, we used the light isotopes ($^{124,126,128}$Xe/$^{130}$Xe) to estimate the contribution of the primordial component relative to the atmosphere $\alpha_{\mathrm{prim/atm}}$

$$\alpha_{\mathrm{prim/atm}}=\frac{({}^{i}\mathrm{Xe}/{}^{130}\mathrm{Xe})_{\mathrm{Eifel}}-({}^{i}\mathrm{Xe}/{}^{130}\mathrm{Xe})_{\mathrm{atm}}}{({}^{i}\mathrm{Xe}/{}^{130}\mathrm{Xe})_{\mathrm{Q}}-({}^{i}\mathrm{Xe}/{}^{130}\mathrm{Xe})_{\mathrm{Eifel}}}$$

where $i=124$, 126 or 128. We used a Monte Carlo method to propagate uncertainties in the isotopic ratios. As an example, the distribution of $\alpha_{\mathrm{prim/atm}}$ obtained using the isotopic ratio $^{124}$Xe/$^{130}$Xe is shown in Extended Data Fig. 2. The average of all $\alpha_{\mathrm{prim/atm}}$ values obtained for $i=124$, 126 and 128 is $16\%\pm2\%$ ($\pm1\sigma$). This value was then used to determine the isotopic composition of a mixture of Q-Xe and atmospheric Xe for the heavy isotopes. The uncertainty in this initial composition was calculated using a Monte Carlo propagation on the uncertainty in $\alpha_{\mathrm{prim/atm}}$ (see the normal distribution in Extended Data Fig. 2).

Second, this initial isotopic composition (renormalized to $^{136}$Xe) was used to compute the relative contributions of the initial component (atmospheric Xe and Q-Xe), $^{\mathrm{Pu}}$Xe and $^{\mathrm{U}}$Xe required to match the isotopic ratios of the Eifel gases ($^{130-132}$Xe/$^{136}$Xe). We used $^{131}$Xe and $^{132}$Xe to constrain the nature of the fissiogenic component because these two isotopes are the most discriminant[31]. The linear system that was solved is

$$\left(\frac{{}^{i}\mathrm{Xe}}{{}^{136}\mathrm{Xe}}\right)_{\mathrm{Eifel}}=\beta_{\mathrm{initial}}\left(\frac{{}^{i}\mathrm{Xe}}{{}^{136}\mathrm{Xe}}\right)_{\mathrm{initial}}+\beta_{\mathrm{Pu_{Xe}}}\left(\frac{{}^{i}\mathrm{Xe}}{{}^{136}\mathrm{Xe}}\right)_{\mathrm{Pu_{Xe}}}+\beta_{\mathrm{U_{Xe}}}\left(\frac{{}^{i}\mathrm{Xe}}{{}^{136}\mathrm{Xe}}\right)_{\mathrm{U_{Xe}}}$$

where $i=130$, 131 or 132 (3 equations), $\left(\frac{{}^{i}\mathrm{Xe}}{{}^{136}\mathrm{Xe}}\right)_{\mathrm{initial}}$ is the initial composition built during the first stage, $\left(\frac{{}^{i}\mathrm{Xe}}{{}^{136}\mathrm{Xe}}\right)_{\mathrm{Pu_{Xe}}}$ and $\left(\frac{{}^{i}\mathrm{Xe}}{{}^{136}\mathrm{Xe}}\right)_{\mathrm{U_{Xe}}}$ are the averages of the fissiogenic spectra from ref. 1, and $\beta_{\mathrm{initial}}$ $\beta_{\mathrm{Pu_{Xe}}}$ and $\beta_{\mathrm{U_{Xe}}}$ are the contributions of each component. We used the same approach as that adopted in ref. 6; that is, we used a MATLAB code with the *lsqlin* function, which minimizes the sum of the squared residuals. Because each component was normalized to the uncertainty in the isotopic composition of the Eifel gas, this sum corresponds to a $\chi^2$ value. We used a Monte Carlo method to propagate the errors in the isotopic composition of the Eifel gas as well as in the initial composition determined during the first stage. A convergence of the results was achieved using $10^5$ simulations. The $\chi^2$ values computed for each simulation are shown in Extended Data Fig. 3. 75% of the $\chi^2$ values are less than 3.

The final results are presented in Extended Data Fig. 4 ($\beta_{\mathrm{initial}}$) and Extended Data Fig. 5 ($\beta_{\mathrm{Pu_{Xe}}}$). The fraction of the initial component (atmospheric Xe and Q-Xe) in the final composition is $97.7\%\pm0.26\%$.

Virtually no simulation leads to a substantial contribution from $^{\mathrm{U}}$Xe; we demonstrate that $\beta_{\mathrm{U_{Xe}}}={}^{136}\mathrm{Xe_U}=0$. To fit a normal distribution to $\beta_{\mathrm{Pu_{Xe}}}$, we had to remove some of the very low values (Extended Data Fig. 5), which resulted in a $^{\mathrm{Pu}}$Xe contribution of $2.26\%\pm0.28\%$.

Because $^{136}\mathrm{Xe_U}=0$, the range for the $^{136}\mathrm{Xe_{Pu}}/(^{136}\mathrm{Xe_{Pu}}+{}^{136}\mathrm{Xe_U})$ ratio is 0.8–1.0 ($\pm1\sigma$). The $^{129}\mathrm{Xe_I}/^{136}\mathrm{Xe_{Pu}}$ ratio was computed using

$$\frac{{}^{129}\mathrm{Xe_I}}{{}^{136}\mathrm{Xe_{Pu}}}=\frac{({}^{129}\mathrm{Xe}/{}^{132}\mathrm{Xe})_{\mathrm{Eifel}}-\beta_{\mathrm{initial}}({}^{129}\mathrm{Xe}/{}^{132}\mathrm{Xe})_{\mathrm{initial}}}{\beta_{\mathrm{Pu_{Xe}}}({}^{136}\mathrm{Xe}/{}^{132}\mathrm{Xe})_{\mathrm{Pu_{Xe}}}}$$

and the errors in $(^{129}\mathrm{Xe}/^{132}\mathrm{Xe})_{\mathrm{Eifel}}$, $\beta_{\mathrm{initial}}$ and $\beta_{\mathrm{Pu_{Xe}}}$ were propagated using the Monte Carlo method. The value obtained for $^{129}\mathrm{Xe_I}/^{136}\mathrm{Xe_{Pu}}$ is $2.1\pm1.6$ ($\pm1\sigma$), which was then used to compute the closure ages of the Eifel gas mantle source regions.

**Code availability.** The code for this Letter is available by contacting G.A. (gavice@crpg.cnrs-nancy.fr).

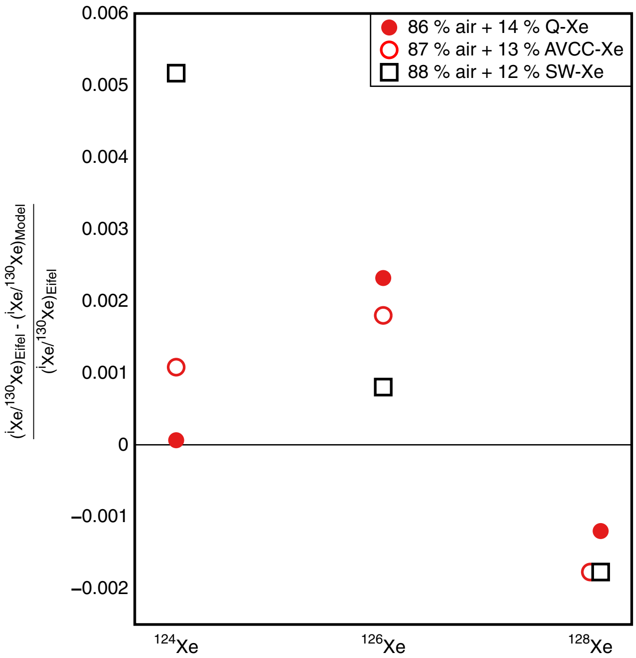
**Sample size.** No statistical methods were used to predetermine sample size.

**Closure ages and I/Pu heterogeneity.** The closure ages (in millions of years) of the Eifel gas mantle source regions were calculated using (see, for example, ref. 12)

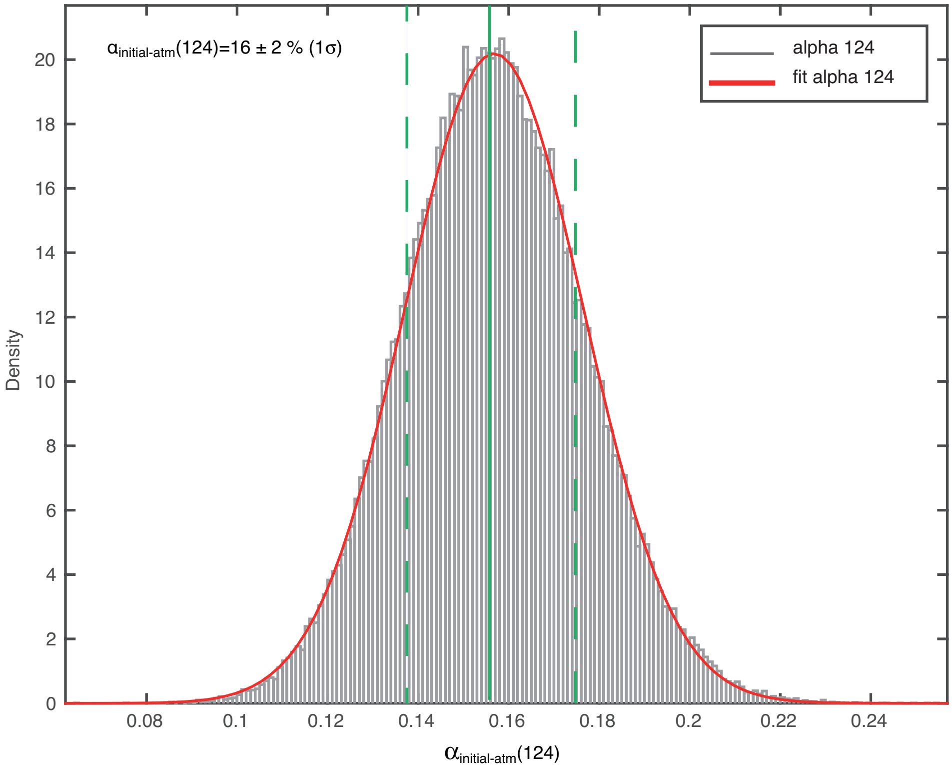
$$t=\frac{1}{\lambda_{244}-\lambda_{129}}\ln\left[\left(\frac{{}^{129}\mathrm{Xe_I}}{{}^{136}\mathrm{Xe_{Pu}}}\right)\left(\frac{{}^{238}\mathrm{U}}{{}^{127}\mathrm{I}}\right)_0\left(\frac{{}^{244}\mathrm{Pu}}{{}^{238}\mathrm{U}}\right)_0\left(\frac{{}^{127}\mathrm{I}}{{}^{129}\mathrm{I}}\right)_0{}^{136}\mathrm{Y}_{244}\right]$$

where $\lambda_{244}=8.45\times10^{-3}\,\mathrm{Myr}^{-1}$ and $\lambda_{129}=4.41\times10^{-2}\,\mathrm{Myr}^{-1}$ are the decay constants of $^{244}$Pu and $^{129}$I, respectively, $^{136}\mathrm{Y}_{244}$ is the yield of fission of $^{244}$Pu for production of $^{136}$Xe (ref. 1), and $^{238}\mathrm{U}_0$, $^{244}\mathrm{Pu}_0$, $^{129}\mathrm{I}_0$ and $^{127}\mathrm{I}_0$ are the initial abundances (in mol) of parent and stable nuclides. Using $\mathrm{U}_0=40$ p.p.b. and $\mathrm{I}_0=6.4$ p.p.b. (ref. 22), we obtained a closure age of $98^{+41}_{-16}$ Myr. This age is relatively insensitive to the initial uranium content ($\mathrm{U}_0$) of the bulk-silicate Earth, whereas the initial iodine content ($\mathrm{I}_0$) is important (see Extended Data Fig. 7 for the sensitivity of the closure age to variable initial I/Pu ratios).

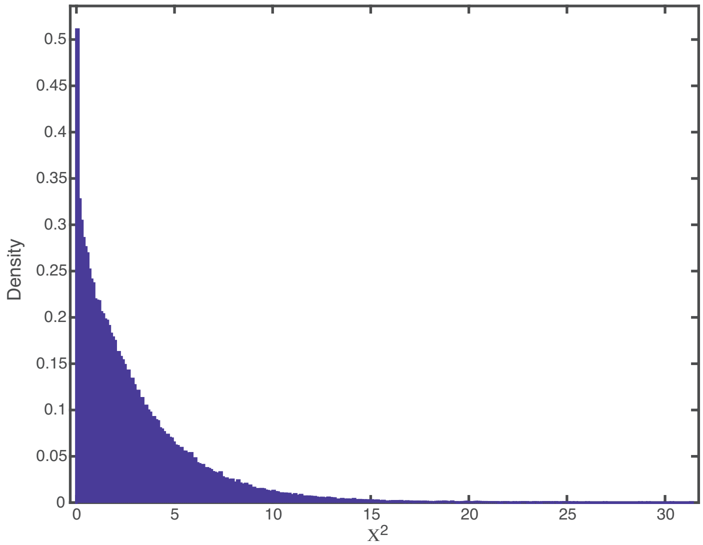
31. Ozima, M. & Podosek, F. A. *Noble Gas Geochemistry* 2nd edn, 22 (Cambridge Univ. Press, 2002).



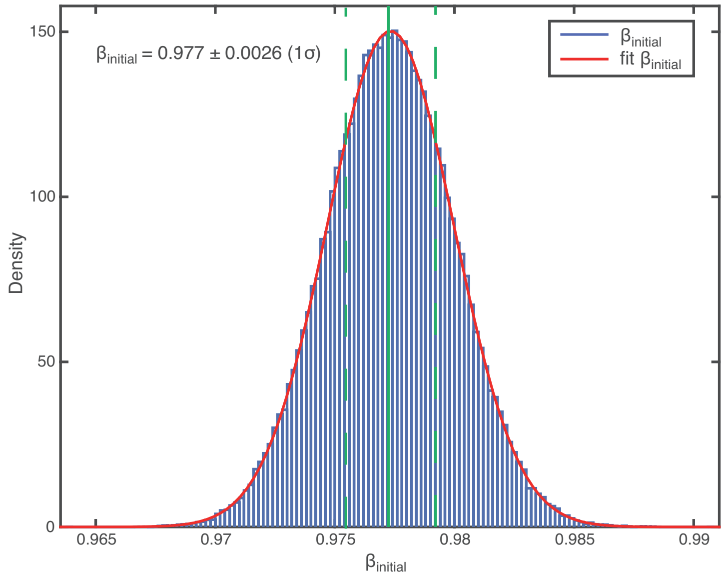

**Extended Data Figure 1 | Residuals of the different mixing possibilities.** Calculations were performed for the light isotopes ($^{124-128}Xe$) using the isotopic compositions of air (typically about 87%) and Q-Xe, AVCC-Xe or SW-Xe (typically about 13%). The best fit is achieved by taking either AVCC-Xe or Q-Xe as the primordial component. SW-Xe does not produce an adequate fit and therefore is not a suitable candidate for this component (as also shown in Fig. 1).



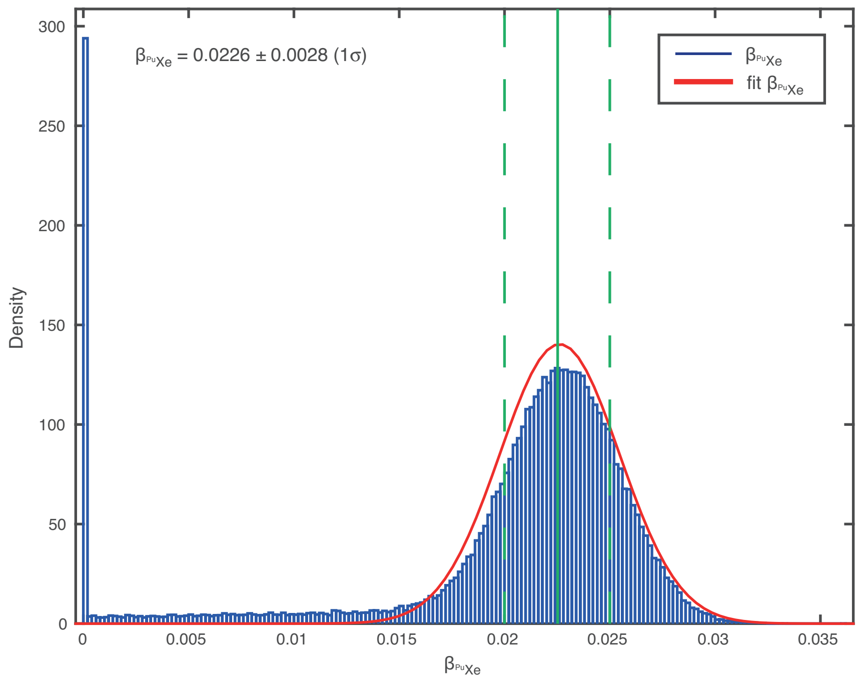

**Extended Data Figure 2 | Deconvolution of the proportion of the primordial component (Q-Xe) relative to the atmosphere for $^{124}Xe/^{130}Xe$.** The red line represents the result of the normal fit. The solid green line depicts the mean value and the dashed green lines depict the error range of $\pm 1\sigma$.



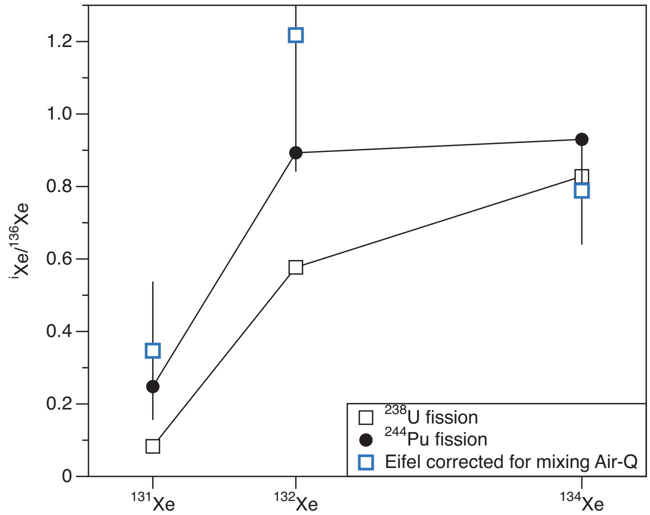

**Extended Data Figure 3 | Range of $\chi^2$ values obtained from the simulations.** Approximately 75% of the values are less than 3.



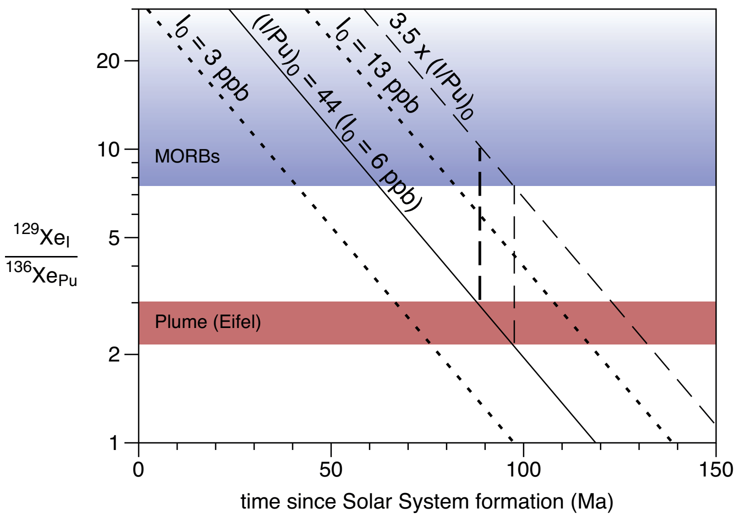

**Extended Data Figure 4 | Fraction of initial component required to fit the isotopic composition of the Eifel gas.** The solid green line depicts the mean value and the dashed green lines depict the error range of $\pm 1\sigma$.




**Extended Data Figure 5 | Fraction of Pu-Xe required to fit the isotopic composition of the Eifel gas.** Some very low values (those less than $10^{-5}$) were excluded from the calculations, resulting in a mean of 2.26% (green line) and a standard deviation of 0.28% ($1\sigma$).




**Extended Data Figure 6 | Isotopic composition of heavy isotopes ($^{131-134}$Xe).** The data are normalized to $^{136}$Xe of the Eifel gas after correction for atmospheric and primitive chondritic contributions, and compared to the fission spectrum of $^{131-136}$Xe produced by spontaneous fission of $^{238}$U and $^{244}$Pu. Excesses in heavy isotopes are compatible with spontaneous fission of $^{244}$Pu.




**Extended Data Figure 7 | Closure ages calculated from the $^{129}Xe_{I}/^{136}Xe_{Pu}$ ratios.** See Methods for details of the computation method. A younger closure age for the upper mantle is achieved only if the I/Pu ratio is at least 3.5 times higher than the lower-mantle source.

**Extended Data Table 1 | Xenon isotopic ratios measured in aliquots of the Eifel gas**

| | $^{130}Xe$ | $^{124}Xe/^{130}Xe$ | $^{126}Xe/^{130}Xe$ | $^{128}Xe/^{130}Xe$ | $^{129}Xe/^{130}Xe$ | $^{131}Xe/^{130}Xe$ | $^{132}Xe/^{130}Xe$ | $^{134}Xe/^{130}Xe$ | $^{136}Xe/^{130}Xe$ |
|---|---|---|---|---|---|---|---|---|---|
| Eif_1 | 7.9E-14 | 0.024225 | 0.022345 | 0.474745 | 6.654409 | 5.215284 | 6.618798 | 2.560233 | 2.187465 |
| | | 0.000109 | 0.000078 | 0.000949 | 0.007304 | 0.005177 | 0.007846 | 0.002519 | 0.002158 |
| Eif_2 | | 0.023910 | 0.021997 | 0.474385 | 6.590819 | 5.159161 | 6.548295 | 2.532569 | 2.168481 |
| | | 0.000130 | 0.000091 | 0.000948 | 0.009207 | 0.006145 | 0.008410 | 0.003489 | 0.002567 |
| Eif_3 | | 0.024502 | 0.022549 | 0.478018 | 6.681859 | 5.221782 | 6.655400 | 2.558332 | 2.245266 |
| | | 0,000100 | 0.000156 | 0.001194 | 0.012001 | 0.008812 | 0.010520 | 0.005035 | 0.004652 |
| Eif_4 | | 0.024168 | 0.022220 | 0.474905 | 6.657195 | 5.198561 | 6.611074 | 2.553088 | 2.168967 |
| | | 0.000291 | 0.000243 | 0.002040 | 0.014614 | 0.011353 | 0.013715 | 0.005527 | 0.006634 |
| Eif_5 | | 0.024139 | 0.022287 | 0.483613 | 6.646030 | 5.204344 | 6.622836 | 2.590335 | 2.191245 |
| | | 0,000073 | 0.000071 | 0.000580 | 0.004642 | 0.004649 | 0.005234 | 0.002039 | 0.001730 |
| Eif_6 | 4.8E-14 | 0.023824 | 0.021997 | 0.473584 | 6.643284 | 5.210670 | 6.574350 | 2.604929 | 2.205901 |
| | | 0.000147 | 0.000093 | 0.000994 | 0.007954 | 0.005690 | 0.008443 | 0.003588 | 0.002612 |
| Eif_7 | | 0.024149 | 0.022200 | 0.478799 | 6.713198 | 5.209149 | 6.602612 | 2.552671 | 2.171623 |
| | | 0.000172 | 0.000126 | 0.001052 | 0.013397 | 0.009307 | 0.013689 | 0.004772 | 0.004285 |
| Eif_8 | | 0.024435 | 0.022229 | 0.477518 | 6.664280 | 5.223837 | 6.607845 | 2.567031 | 2.193414 |
| | | 0.000120 | 0.000129 | 0.000954 | 0.008644 | 0.006741 | 0.007181 | 0.003031 | 0.003246 |
| Eif_9 | | 0.023748 | 0.022094 | 0.476777 | 6.672970 | 5.199317 | 6.610295 | 2.587957 | 2.206022 |
| | | 0.000309 | 0.000283 | 0.002715 | 0.021306 | 0.014451 | 0.019591 | 0.007130 | 0.007618 |
| Eif_10 | | 0.023691 | 0.022016 | 0.474975 | 6.597032 | 5.166969 | 6.544023 | 2.538413 | 2.165106 |
| | | 0.000134 | 0.000089 | 0.001092 | 0.009874 | 0.008206 | 0.011637 | 0.003996 | 0.002991 |
| Eif_11 | 3.2E-14 | 0.024006 | 0.022239 | 0.481571 | 6.667457 | 5.230183 | 6.637291 | 2.566452 | 2.178160 |
| | | 0.000254 | 0.000232 | 0.001588 | 0.011310 | 0.008307 | 0.010491 | 0.005051 | 0.004298 |
| Eif_12 | | 0.023815 | 0.022016 | 0.475436 | 6.657245 | 5.221550 | 6.583217 | 2.569755 | 2.197417 |
| | | 0.000130 | 0.000170 | 0.001235 | 0.009964 | 0.008811 | 0.010406 | 0.004551 | 0.003469 |
| Eif_13 | | 0.023881 | 0.022278 | 0.477177 | 6.673390 | 5.224593 | 6.648375 | 2.570202 | 2.197580 |
| | | 0.000123 | 0.000127 | 0.000715 | 0.009322 | 0.006742 | 0.006568 | 0.004046 | 0.003036 |
| Eif_14 | | 0.023729 | 0.021977 | 0.477007 | 6.665152 | 5.215294 | 6.600769 | 2.594237 | 2.190850 |
| | | 0.000152 | 0.000141 | 0.001239 | 0.009311 | 0.007765 | 0.007173 | 0.004340 | 0.003242 |
| Eif_15 | | 0.023939 | 0.022462 | 0.475946 | 6.648185 | 5.208323 | 6.582498 | 2.562519 | 2.188883 |
| | | 0.000133 | 0.000158 | 0.000761 | 0.008624 | 0.006204 | 0.008454 | 0.003026 | 0.003239 |
| Eifel | | 0.024011 | 0.022194 | 0.476964 | 6.655500 | 5.207268 | 6.603179 | 2.567248 | 2.190425 |
| | | 0.000065 | 0.000046 | 0.000707 | 0.007827 | 0.005231 | 0.008213 | 0.005221 | 0.005209 |

Individual Xe isotopic ratio measurements ('Eif_*n*') for three different aliquots of the same gas. Errors ($1\sigma$) are reported in the cell below each ratio. The 'Eifel' ratios represent the average of 15 measurements ('Eif_1' to 'Eif_15') of three aliquots of the same gas. The amounts of $^{130}Xe$ are in mol and errors are 5%.

# LETTER

doi:10.1038/nature17398

# The genetic program for cartilage development has deep homology within Bilateria

Oscar A. Tarazona[1,2], Leslie A. Slota[3], Davys H. Lopez[3], GuangJun Zhang[2]† & Martin J. Cohn[1,2,3]

**The evolution of novel cell types led to the emergence of new tissues and organs during the diversification of animals[1]. The origin of the chondrocyte, the cell type that synthesizes cartilage matrix, was central to the evolution of the vertebrate endoskeleton. Cartilage-like tissues also exist outside the vertebrates, although their relationship to vertebrate cartilage is enigmatic. Here we show that protostome and deuterostome cartilage share structural and chemical properties, and that the mechanisms of cartilage development are extensively conserved—from induction of chondroprogenitor cells by Hedgehog and β-catenin signalling, to chondrocyte differentiation and matrix synthesis by SoxE and SoxD regulation of clade A fibrillar collagen (*ColA*) genes—suggesting that the chondrogenic gene regulatory network evolved in the common ancestor of Bilateria. These results reveal deep homology of the genetic program for cartilage development in Bilateria and suggest that activation of this ancient core chondrogenic network underlies the parallel evolution of cartilage tissues in Ecdysozoa, Lophotrochozoa and Deuterostomia.**

Cartilage forms the embryonic endoskeleton of all vertebrates and has been widely considered to be a vertebrate-specific tissue[2,3]. This endoskeletal connective tissue is formed by non-adjacent cells embedded in abundant extracellular matrix (ECM) that is rich in collagen and acidic glycosaminoglycans (GAGs)[4,5]. Cartilage-like tissues have been recognized in invertebrate species scattered throughout the Protostomia (Fig. 1a); however, the relationship of these tissues to the bona fide cellular cartilage of vertebrates has long been debated[4,6,7] and the evolutionary origin of cartilage and its parent cell type, the chondrocyte, is unknown[4,8–10].

To determine whether invertebrate cartilage-like tissues have structural and/or chemical similarities to vertebrate cartilage, we compared the structure and matrix composition of these tissues in adults of two distantly related protostomes, the cuttlefish *Sepia bandensis* from the Lophotrochozoa and the horseshoe crab *Limulus polyphemus* from the Ecdysozoa, which are among the best known examples of invertebrate cartilage-like tissues[4,11]. Cartilage-like tissues in both species are composed of cells embedded in abundant ECM rich in collagen and acidic GAGs, and form conspicuous endoskeletal structures (Figs 1b–g and 2a–c, m, n and Supplementary Video 1). We then investigated chondrogenesis in *Sepia* and *Limulus*. In both species, cartilage development begins during late stages of organogenesis with the formation of pre-chondrogenic mesenchymal cell condensations that later secrete an ECM rich in collagen and acidic GAGs, mirroring the process of vertebrate chondrogenesis (Extended Data Fig. 1).

To test whether a common genetic program for cartilage development is conserved across the Bilateria, we asked whether *Sepia* and *Limulus* cartilages express pro-orthologues of the vertebrate collagen2α1 (*Col2α1*) gene, which encodes type II collagen, the most abundant protein in vertebrate cartilage ECM. We isolated two clade A collagen (*ColA*) genes from *Sepia* and one from *Limulus*

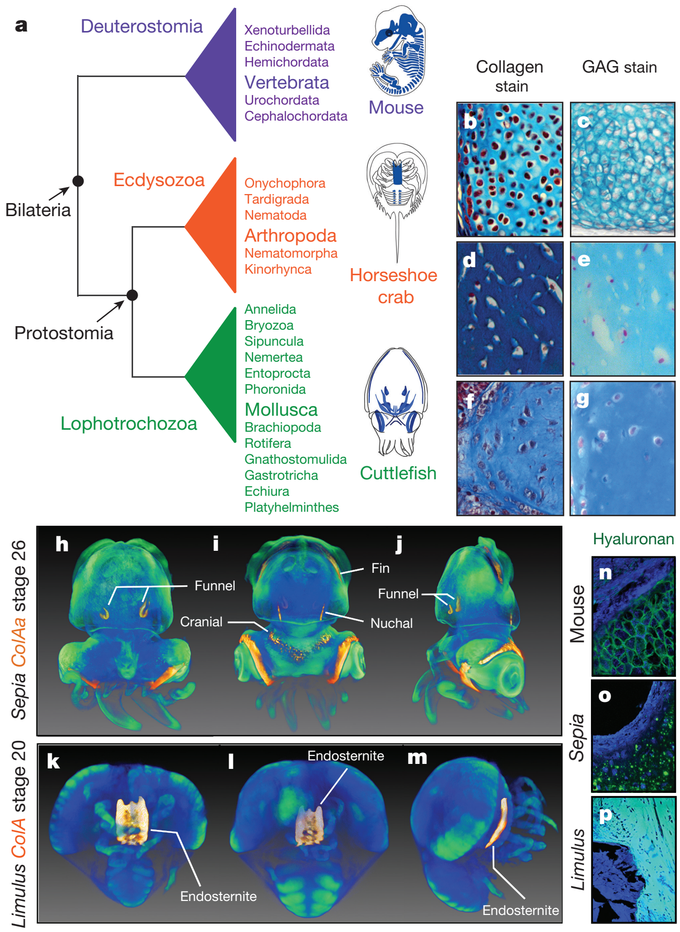



**Figure 1 | Protostome invertebrate cartilage is structurally similar to vertebrate cartilage. a**, Cartilage has evolved in Deuterostomia, Ecdysozoa and Lophotrochozoa. Cartilaginous endoskeletons of mouse, *Limulus* and *Sepia* are shown in blue (see also Supplementary Video 1). **b**–**g**, Sections through cartilage of mouse vertebra (**b**, **c**), horseshoe crab endosternite (**d**, **e**) and cuttlefish funnel (**f**, **g**). Masson's trichrome stain shows high content of collagen (**b**, **d**, **f**) and alcian blue stain shows high content of acidic GAGs (**c**, **e**, **g**) in all three cartilages. **h**–**m**, *ColA* expression in *Sepia* (**h**–**j**) and *Limulus* (**k**–**m**) embryos scanned with optical projection tomography (see Supplementary Videos 2 and 3). **h**–**j**, *ColAa* transcripts in *Sepia* embryos localize to numerous cartilages during chondrogenesis. **k**–**m**, *ColA* transcripts in the endosternite cartilage of *Limulus* embryos. Embryos shown in ventral (**h**, **k**), dorsal (**i**, **l**), and lateral (**j**, **m**) orientations. **n**–**p**, Cartilage ECM in vertebrates and protostomes is positive for hyaluronan.

[1]Howard Hughes Medical Institute, UF Genetics Institute, University of Florida, PO Box 103610, Gainesville, Florida 32610, USA. [2]Department of Biology, University of Florida, PO Box 103610, Gainesville, Florida 32610, USA. [3]Department of Molecular Genetics and Microbiology, University of Florida College of Medicine, PO Box 103610, Gainesville, Florida 32610, USA.
†Present address: Department of Comparative Pathobiology, Purdue University College of Veterinary Medicine, 625 Harrison Street, West Lafayette, Indiana 47907, USA.

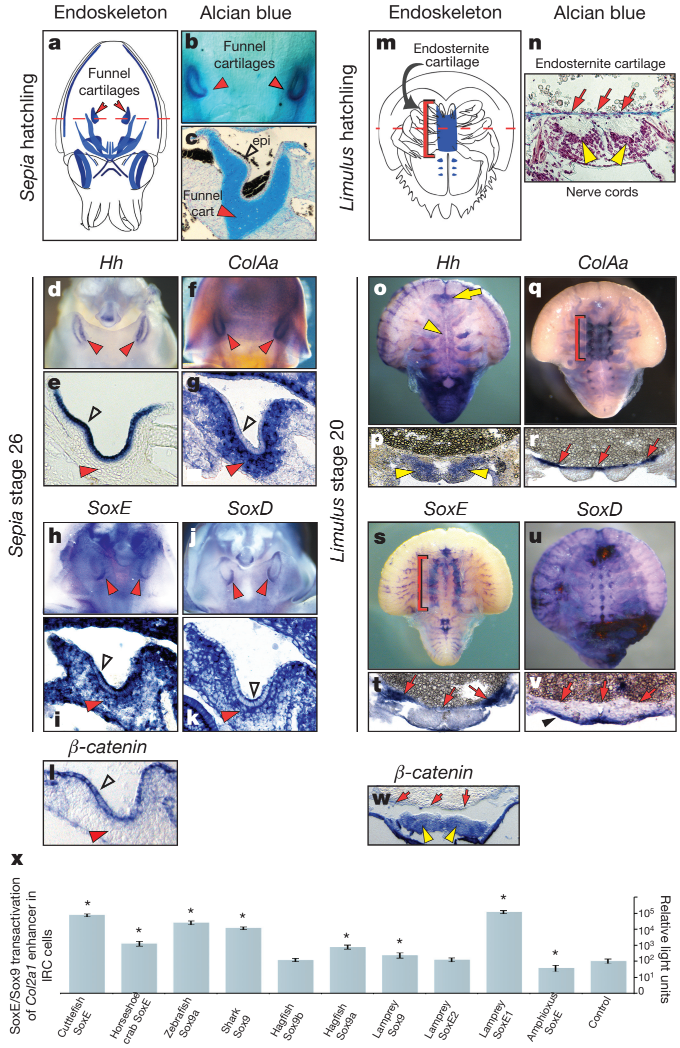



**Figure 2 | Deep conservation of gene expression during protostome cartilage development. a**, The *Sepia* hatchling endoskeleton. Dashed line indicates the plane of sections at the funnel cartilages (red arrowheads). **b**–**c**, Funnel cartilage in hatchlings stained with alcian blue in whole-mount (**b**) and in section (**c**). Cartilage, red arrowhead; funnel epithelium, black open arrowhead. **d**–**l**, Gene expression in *Sepia* embryos during chondrogenesis of funnel cartilage. Cartilage precursors in whole-mounts (**d**, **f**, **h** and **j**) and on sections (**e**, **g**, **i**, **k** and **l**), red arrowheads; funnel epithelium (**e**, **g**, **i**, **k** and **l**), open arrowheads. **m**, The *Limulus* hatchling endoskeleton. Dashed line indicates the plane of sections at the endosternite cartilage (red bracket). **n**, Endosternite cartilage (red arrows) of *Limulus* hatchling stained with alcian blue, located dorsal to the two nerve cords (yellow arrowheads). **o**–**w**, Gene expression during endosternite chondrogenesis in *Limulus* embryos. Nerve cords in whole-mounts (**o**) and on sections (**p**, **w**), yellow arrowheads; brain (**o**), yellow arrow. Pre-chondrogenic domains of endosternite chondrogenesis in whole-mounts (**o**, **q**, **s** and **u**), red brackets; and on sections (**p**, **r**, **t**, **v**, **w**), red arrows; ectoderm (**v**), black arrowhead. **x**, Luciferase reporter assay of SoxE/Sox9 transactivation of the human *COL2A1* enhancer in IRC cells. Significant differences over control (two-tailed *t*-test; $P < 0.05$; $n = 4$), asterisks; error bars, s.d. Each luciferase experiment was repeated four times, with four replicates per experiment.

(Extended Data Fig. 2) and analysed their expression during cartilage formation in both species. In *Sepia* embryos at stage 26, *ColAa* and *ColAb* expression localized to numerous regions of chondrogenesis, including the funnel, nuchal, fin and cranial cartilages (Fig. 1h–j, Extended Data Fig. 3 and Supplementary Video 2). Similarly, in *Limulus* embryos at stage 20, we detected *ColA* expression in the developing endosternite (Fig. 1k–m and Supplementary Video 3) and gill cartilages (Fig. 2q).

Hyaluronan, a non-sulfated GAG, is an essential component of vertebrate cartilage[12,13]. Although largely absent outside vertebrates, purification from a mollusc[14] led us to screen for hyaluronan in invertebrate cartilage. Using a hyaluronan binding peptide, we determined that *Sepia* and *Limulus* cartilage ECM is positive for hyaluronan (Fig. 1n–p). Hyaluronan synthases, however, could not be identified in protostome transcriptomic and genomic databases, suggesting that (1) hyaluronan could be synthesized by alternative mechanisms, (2) hyaluronan synthases could be present but unidentified in these groups or (3) hyaluronan synthases evolved multiple times, perhaps from chitin synthases, as proposed for the origin of vertebrate hyaluronan synthases[15]. The discovery of *ColA* and hyaluronan in these tissues reveals that key structural molecular components of cartilage are shared between invertebrate and vertebrate cartilages, and suggests that invertebrate cartilage is fibrillar-collagen-based.

We next investigated whether cell signalling proteins and transcription factors that function upstream of *Col2α1* in vertebrate chondrogenesis also are expressed during invertebrate chondrogenesis. The Hedgehog signalling pathway plays essential roles in early vertebrate chondrogenesis, where Shh regulates transcriptional activation of *Sox5/6/9* and Ihh regulates cartilage proliferation and differentiation[16–18]. We cloned *Hh* from *Sepia*, and analysis of its expression revealed that cartilage differentiation takes place in close proximity to *Hh* expression domains (Fig. 2d, e and Extended Data Fig. 4). In *Sepia* embryos at stage 26, *ColAa* and *ColAb* are expressed in a 'U-shaped' domain of mesenchymal pre-cartilaginous cells immediately adjacent to the *Hh*-expressing epithelium, in the region that will later form the funnel cartilage (Fig. 2d–g and Extended Data Fig. 4a, b).

Downstream of Hedgehog signalling, vertebrate Sox9, Sox5 and Sox6 function as master regulators of chondrogenesis by directly activating transcription of *Col2α1* (refs 19–21). We isolated invertebrate pro-orthologues of *Sox9* and *Sox5*/*Sox6*—*SoxE* and *SoxD*, respectively (Extended Data Fig. 2)—and found that *Sepia* funnel cartilage condensations express both *SoxE* (Fig. 2h, i and Extended Data Fig. 5) and *SoxD* (Fig. 2j, k and Extended Data Fig. 5), mirroring the demarcation of vertebrate cartilage condensations by *Sox9* and *Sox5/6* (ref. 20). Thus, as in vertebrates, *SoxE*, *SoxD* and *ColA* are co-expressed in the developing cartilaginous skeleton of *Sepia* embryos (Fig. 2f–k).

Sox9 and β-catenin have opposing functions in vertebrate chondrogenesis: they inhibit each other's transcriptional activity and β-catenin functions as an anti-chondrogenic transcriptional regulator[22–24]. Indeed, reduction of *β-catenin* mRNA and protein levels in *Sox9*-positive chondroprogenitor cells is necessary for cartilage differentiation[22–24]. To test whether this regulatory relationship is conserved in *Sepia* cartilage, we cloned *β-catenin* and found that it is expressed in funnel mesenchyme and in the overlying epithelium at stages 24–25 (Extended Data Fig. 5e), but by stage 26, *β-catenin* mRNA is undetectable in pre-cartilaginous condensations (Fig. 2l). As in vertebrates, downregulation of *β-catenin* in *Sepia* chondroprogenitors precedes the onset of differentiation (Extended Data Fig. 1).

To test whether conservation of a core chondrogenic network extends to Ecdysozoa, we cloned and analysed expression of *Hh*, *SoxE*, *SoxD* and *β-catenin* during *Limulus* chondrogenesis. In *Limulus* embryos, the endosternite cartilage forms immediately adjacent to the *Hh*-expressing ventral nerve cords (Fig. 2o–p and Extended Data Fig. 1). The pre-chondrogenic condensation of the endosternite can be identified as a thin plate of *ColA*-expressing cells dorsal to the paired nerve cords (Fig. 2q, r). Similarly, *SoxE* is expressed throughout the developing endosternite plate (Fig. 2s, t) and gill cartilages (Extended Data Fig. 6c), although *SoxD* was not detectable in these tissues (Fig. 2u–v). In *Limulus*, as in *Sepia* and vertebrates, *β-catenin* is downregulated in *SoxE*/*ColA*-expressing cells before cartilage differentiation (Fig. 2w). Taken together, our analysis of *Sepia* and *Limulus* chondrogenesis indicates that the network of structural and regulatory genes required for vertebrate cartilage development has deeply conserved patterns of expression in the three major lineages of Bilateria.

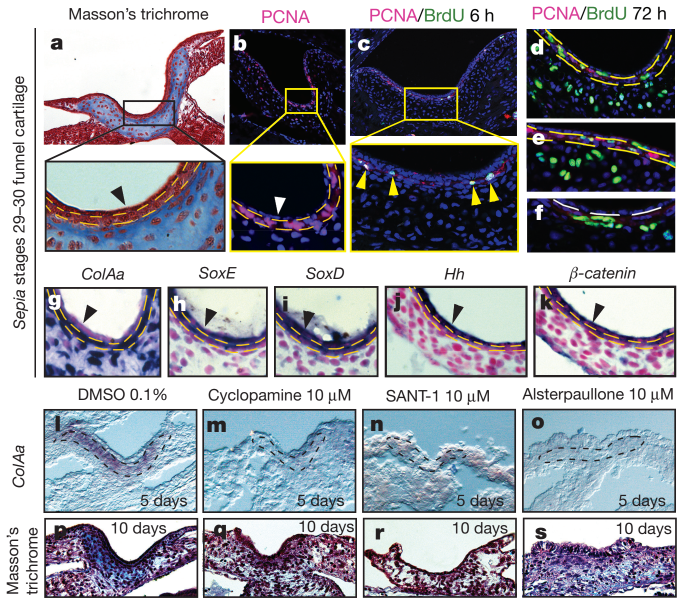



**Figure 3 | Cuttlefish chondrogenesis is regulated positively by *Hh* signalling and negatively by *β-catenin*. a**, At stages 29–30 a non-cartilaginous, undifferentiated cell layer (yellow broken line outline) lies below the funnel epithelium (black arrowhead). **b**, Proliferating cell nuclear antigen (PCNA) immunoreactivity indicates cell proliferation is restricted to this undifferentiated cell layer (between yellow broken lines); funnel epithelium, white arrowhead. **c–f**, Pulse-chase assay of BrdU incorporation at (**c**) 6 h exposure and (**d–f**) 72 h of incubation after initial exposure. At 6 h, PCNA/BrdU are co-localized to the proliferative zone (yellow arrowheads). After 72 h, labelled cells advance into the funnel cartilage (**d–e**) and also contribute to the cell pool in the proliferative zone (**f**). Broken white line marks basal lamina of funnel epithelium. **g–k**, The proliferative zone expresses *ColAa* (**g**), *SoxE* (**h**), *SoxD* (**i**), and *β-catenin* (**k**); the funnel epithelium expresses *Hh* (**j**) and *β-catenin* (**k**). **l–o**, *ColAa* expression in the funnel region after 5 days of treatment with cyclopamine, SANT-1 and alsterpaullone ($n = 4$ embryos each treatment). In DMSO controls, pre-cartilaginous mesenchyme (outlined by dashed lines) expresses *ColAa* (**l**), whereas *ColAa* is undetectable in the funnel of embryos treated with (**m**) cyclopamine, (**n**) SANT-1 or (**o**) alsterpaullone. **p–s**, Masson's trichrome staining of funnel region after 10 days of treatment ($n = 4$ embryos per treatment) shows that DMSO control embryos undergo cartilage differentiation (**t**), but embryos treated with cyclopamine (**q**), SANT-1 (**r**) or alsterpaullone (**s**) do not differentiate into cartilage and lack a collagenous matrix (see Supplementary Table 4 for number of embryos treated and analysed).

To determine whether invertebrate SoxE proteins could function as transcriptional regulators of *ColA* genes, we tested the ability of *Sepia* and *Limulus* SoxE proteins to activate the human *COL2A1* cartilage-specific enhancer[25] using a luciferase reporter assay in NIH3T3 and immortalized rat chondrocyte (IRC) cells. Quantification of luciferase activity revealed significant transactivation by *Sepia* and *Limulus* SoxE, with efficiencies equal to or greater than Sox9/SoxE from lamprey, hagfish, shark and zebrafish (Fig. 2x and Extended Data Fig. 5). Thus, *Sepia* and *Limulus* SoxE proteins have transactivation functions similar to vertebrate Sox9 proteins.

Vertebrate cartilage growth occurs by progression of chondrocytes from resting zones (at the epiphyseal ends) into the zones of proliferation, maturation and hypertrophy[26]. Prehypertrophic cells secrete Ihh, which regulates proliferation and differentiation of adjacent chondrocytes[26]. In *Sepia* embryos at stages 29–30, we observed an undifferentiated layer of proliferating cells between the *Hh*-expressing funnel epithelium and the overtly differentiated chondrocytes (Fig. 3a, b). To test whether *Sepia* funnel cartilage growth involves directional proliferation and maturation, similar to vertebrate cartilage, we performed a pulse-chase 5-bromo-2′-deoxyuridine (BrdU) assay. Chondroprogenitor cells labelled in the proliferative zone (Fig. 3c) were later found deep in the funnel cartilage (Fig. 3d–f), indicating directional growth. Restriction of proliferating cells to the perimeter of the funnel cartilage, adjacent to *Hh*-expressing cells, and expression of *SoxE*, *SoxD*, *β-catenin* and *ColAa*/*ColAb* in this highly-proliferative undifferentiated layer suggest that appositional growth of the *Sepia* funnel cartilage occurs at the ends of the element, reminiscent of the growth pattern of vertebrate cartilage (Fig. 4a).

We then tested whether expression of *Hh* and degradation of β-catenin are necessary for *ColA* expression and differentiation of funnel cartilage in *Sepia*, as is the case in vertebrates[22–26]. We used the small molecules cyclopamine and SANT-1 (inhibitors of Smoothened) to block Hh signalling[27] and alsterpaullone (inhibitor of GSK-3β) to stabilize β-catenin[28]. Five-day treatments were initiated at stages 23–24, before formation of pre-cartilaginous cell condensations but after the appearance of the funnel epithelium and associated mesenchyme. Hh antagonism or β-catenin stabilization resulted in loss of *ColAa* expression in funnel chondroprogenitor cells by stage 26 (Fig. 3l–o). Although *ColA* expression was not maintained, funnel chondroprogenitors continued to proliferate and express *β-catenin* and funnel epithelium continued to express *Hh* (Extended Data Figs 7, 8 and 9n, o), indicating that loss of *ColAa* was not due to toxicity or global effects on transcription.

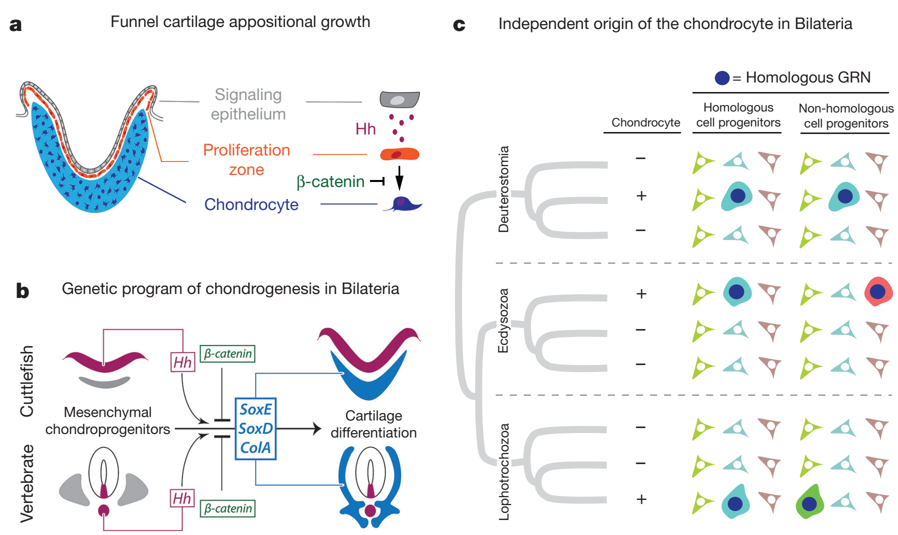



**Figure 4 | Bilaterian cartilage development and the origin of the chondrocyte. a**, Model of cuttlefish funnel cartilage appositional growth. New cartilage is derived from a proliferative layer of chondrocyte precursors. **b**, Conservation of the developmental genetic program of cartilage development between vertebrates and invertebrates. Vertebrate cartilage represented by a mouse vertebra and invertebrate cartilage by a *Sepia* funnel cartilage. **c**, Two hypotheses for independent evolution of cartilage are presented. The homologous cell progenitors model depicts the independent origin of the chondrocyte through the recruitment of a homologous gene regulatory network (GRN; dark blue circle) by the same homologous progenitor cell type (blue fibroblasts) across Bilateria. Alternatively, the non-homologous cell progenitors model depicts independent evolution by activation of the same homologous gene regulatory network but in different, non-homologous progenitor cell types (blue, red and green fibroblasts). See Supplementary Discussion for further details.

Prolonged inhibition of *ColA* in *Sepia* prevented differentiation of cartilage tissue. When embryos were cultured for 10 days, to stage 28, in the presence of Smo antagonists (cyclopamine or SANT-1) or GSK-3β inhibitors (alsterpaullone or BIO), funnel cartilage differentiation was inhibited (Fig. 3q–s and Extended Data Fig. 9a). By contrast, embryos treated with dimethylsulfoxide (DMSO) alone (controls) or with the β-catenin signalling repressors IWR-1 or PNU underwent normal funnel cartilage differentiation, including generation of a conspicuous cartilage ECM (Fig. 3p and Extended Data Fig. 9b, c). Taken together, our finding that Hh and β-catenin signalling have opposite effects on chondrogenesis both in cuttlefish and in vertebrates demonstrates deep conservation of the genetic program for chondrogenesis in Bilateria (Fig. 4b)

Although a single origin of the bilaterian chondrocyte is still plausible, we posit two hypotheses that can account for independent origins of the chondrocyte (Fig. 4c and Supplementary Discussion). In the first scenario, the chondrocyte evolved in parallel, but from a homologous fibroblast-like cell type, in different lineages of Bilateria by recruitment of a homologous gene regulatory network (Fig. 4c). One possible chondroid precursor cell that could have been recruited in parallel in Bilateria is the mesodermal midline cell type that gives rise to the deuterostome notochord and the protostome axochord[29]. However, a variety of cell/tissue types in invertebrate deuterostomes activate components of the chondrogenic gene regulatory network during histogenesis of non-cartilaginous endoskeletal tissues[10,30]. Therefore, an alternative hypothesis is that chondrocytes evolved in parallel by activation of the same gene regulatory network but in non-homologous cell types across the Bilateria (Fig. 4c). Taken together, our data suggest that the core kernel of the chondrogenic gene network that orchestrates cartilage development was probably present in the urbilaterian ancestor and may have been involved in the production of a specialized ECM type. Finally, our results raise the potential for the emergence of invertebrates as new model systems for the study of chondrogenesis, cartilage physiology and regeneration.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 18 July 2014; accepted 4 February 2016.**
**Published online 25 April 2016.**

1. Arendt, D. The evolution of cell types in animals: emerging principles from molecular studies. *Nature Rev. Genet.* **9,** 868–882 (2008).
2. Gans, C. & Northcutt, R. G. Neural crest and the origin of vertebrates: a new head. *Science* **220,** 268–273 (1983).
3. Meulemans, D. & Bronner-Fraser, M. Insights from amphioxus into the evolution of vertebrate cartilage. *PLoS ONE* **2,** e787 (2007).
4. Person, P. & Philpott, D. E. The nature and significance of invertebrate cartilages. *Biol. Rev. Camb. Phil. Soc.* **44,** 1–16 (1969).
5. Zhang, G., Eames, B. F. & Cohn, M. J. Chapter 2. Evolution of vertebrate cartilage development. *Curr. Top. Dev. Biol.* **86,** 15–42 (2009).
6. Schaffer, J. in *Handbuch der mikroskopischen Anatomie des Menschen* Vol. 2 (2) (ed. W von Möllendorf) 1–390 (Springer, 1930).
7. Cowden, R. R. A histochemical study of chondroid tissue in *Limulus* and *Octopus*. *Histochemie* **9,** 149–163 (1967).
8. Hall, B. K. & Gillis, J. A. Incremental evolution of the neural crest, neural crest cells and neural crest-derived skeletal tissues. *J. Anat.* **222,** 19–31 (2013).
9. Cole, A. G. & Hall, B. K. Cartilage is a metazoan tissue; integrating data from nonvertebrate sources. *Acta Zool.* **85,** 69–80 (2004).
10. Zhang, G. & Cohn, M. J. Hagfish and lancelet fibrillar collagens reveal that type II collagen-based cartilage evolved in stem vertebrates. *Proc. Natl Acad. Sci. USA* **103,** 16829–16833 (2006).
11. Cole, A. G. & Hall, B. K. The nature and significance of invertebrate cartilages revisited: distribution and histology of cartilage and cartilage-like tissues within the Metazoa. *Zoology* **107,** 261–273 (2004).
12. Roughley, P. J. The structure and function of cartilage proteoglycans. *Eur. Cell. Mater.* **12,** 92–101 (2006).
13. Matsumoto, K. *et al.* Conditional inactivation of Has2 reveals a crucial role for hyaluronan in skeletal growth, patterning, chondrocyte maturation and joint formation in the developing limb. *Development* **136,** 2825–2835 (2009).
14. Volpi, N. & Maccari, F. Purification and characterization of hyaluronic acid from the mollusc bivalve *Mytilus galloprovincialis*. *Biochimie* **85,** 619–625 (2003).
15. DeAngelis, P. L. Evolution of glycosaminoglycans and their glycosyltransferases: implications for the extracellular matrices of animals and the capsules of pathogenic bacteria. *Anat. Rec.* **268,** 317–326 (2002).
16. Zeng, L., Kempf, H., Murtaugh, L. C., Sato, M. E. & Lassar, A. B. Shh establishes an Nkx3.2/Sox9 autoregulatory loop that is maintained by BMP signals to induce somitic chondrogenesis. *Genes Dev.* **16,** 1990–2005 (2002).
17. Abzhanov, A. & Tabin, C. J. Shh and Fgf8 act synergistically to drive cartilage outgrowth during cranial development. *Dev. Biol.* **273,** 134–148 (2004).
18. Kronenberg, H. M. Developmental regulation of the growth plate. *Nature* **423,** 332–336 (2003).
19. Akiyama, H., Chaboissier, M. C., Martin, J. F., Schedl, A. & de Crombrugghe, B. The transcription factor Sox9 has essential roles in successive steps of the chondrocyte differentiation pathway and is required for expression of Sox5 and Sox6. *Genes Dev.* **16,** 2813–2828 (2002).
20. Smits, P. *et al.* The transcription factors L-Sox5 and Sox6 are essential for cartilage formation. *Dev. Cell* **1,** 277–290 (2001).
21. Lefebvre, V., Li, P. & de Crombrugghe, B. A new long form of Sox5 (L-Sox5), Sox6 and Sox9 are coexpressed in chondrogenesis and cooperatively activate the type II collagen gene. *EMBO J.* **17,** 5718–5733 (1998).
22. Hill, T. P., Später, D., Taketo, M. M., Birchmeier, W. & Hartmann, C. Canonical Wnt/β-catenin signaling prevents osteoblasts from differentiating into chondrocytes. *Dev. Cell* **8,** 727–738 (2005).
23. Day, T. F., Guo, X., Garrett-Beal, L. & Yang, Y. Wnt/β-catenin signaling in mesenchymal progenitors controls osteoblast and chondrocyte differentiation during vertebrate skeletogenesis. *Dev. Cell* **8,** 739–750 (2005).
24. Akiyama, H. *et al.* Interactions between Sox9 and β-catenin control chondrocyte differentiation. *Genes Dev.* **18,** 1072–1087 (2004).
25. Lefebvre, V. *et al.* A 47-bp sequence of the first intron of the mouse proα1(II) collagen gene is sufficient to direct chondrocyte expression. *Ann. NY Acad. Sci.* **785,** 284–287 (1996).
26. St-Jacques, B., Hammerschmidt, M. & McMahon, A. P. Indian hedgehog signaling regulates proliferation and differentiation of chondrocytes and is essential for bone formation. *Genes Dev.* **13,** 2072–2086 (1999).
27. Grimaldi, A. *et al.* A hedgehog homolog is involved in muscle formation and organization of *Sepia officinalis* (Mollusca) mantle. *Dev. Dynam.* **237,** 659–671 (2008).
28. Broun, M., Gee, L., Reinhardt, B. & Bode, H. R. Formation of the head organizer in hydra involves the canonical Wnt pathway. *Development* **132,** 2907–2916 (2005).
29. Lauri, A. *et al.* Development of the annelid axochord: insights into notochord evolution. *Science* **345,** 1365–1368 (2014).
30. Rychel, A. L. & Swalla, B. J. Development and evolution of chordate cartilage. *J. Exp. Zool. B* **308,** 325–335 (2007).

**Supplementary Information** is available in the online version of the paper.

**Author Contributions** O.A.T. and M.J.C. designed the experiments, analysed the data and wrote the paper. O.A.T., L.A.S. and D.H.L. cloned *Sepia* and *Limulus* genes, analysed gene and protein expression, and performed histological analysis of adult tissues. O.A.T. isolated the full-length *ColA* cDNA, performed the optical projection tomography scanning and three-dimensional reconstructions, BrdU and small-molecule treatments, and the experiments on treated and control embryos. G.Z. isolated the full-length *Sox9/SoxE* genes and prepared the expression constructs for amphioxus, lamprey, hagfish and shark, and O.A.T. for zebrafish, cuttlefish and horseshoe crab. O.A.T. performed the cell culture and luciferase assays.

**Acknowledgements** We thank B. Battelle and members of H. J. Brockmann's and David Julian's laboratories for *Limulus* eggs, N. Brown for sharing protocols and reagents, N. Patel and the *Limulus* genome consortium for access to sequence data, and M. Welten and F. Leal for assisting with optical projection tomography and luciferase assays, respectively. This project was supported by the Howard Hughes Medical Institute (to M.J.C.). O.A.T. was a Howard Hughes Medical Institute International Student Research Fellow.

**Author Information** Sequences generated in this work have been deposited in GenBank under accession numbers KP322116–KP322126. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.J.C. (mjcohn@ufl.edu).

## METHODS

No statistical methods were used to predetermine sample size. Embryos were randomized in each experiment. The investigators were not blinded to allocation during experiments and outcome assessment.

**Embryo collection and preparation.** *Sepia phraonis* eggs were obtained from the National Resource Center for Cephalopods, Galveston, Texas, USA. *S. bandensis* and *Sepia officinalis* eggs were purchased from commercial suppliers. Upon arrival at our institution, eggs were cultured in artificial seawater (Petco Real Ocean Water) at 22 °C. Embryos were collected by manual removal of egg cases and were staged according to ref. 31. Embryos used for *in situ* hybridization (ISH) and immunohistochemistry were fixed and processed as previously described[27]. *Limulus* embryos were provided by B. Battelle and members of H. J. Brockmann's and D. Julian's laboratories and were staged according to ref. 32 and processed for ISH as previously described[33].

**Alcian blue and Masson's trichrome stain.** Cartilage was stained with alcian blue staining to reveal GAGs. In vertebrate cartilage, alcian blue can detect highly anionic GAGs, such as hyaluronan and sulfated GAGs. Early biochemical analyses of cartilaginous tissues in cephalopods and horseshoe crabs suggested the presence of highly sulfated GAGs, such as chondroitin sulfate[34–36]. To detect GAGs in *Sepia* and *Limulus* cartilage, we used alcian blue/nuclear fast red staining on paraffin sections. Deparaffinized sections were stained for 30 min in 1% alcian blue (in 3% acetic acid) and counter stained with nuclear fast red for 5 min. Masson's trichrome staining was performed on paraffin sections using a Masson's Trichrome Kit (22-110-648, Richard-Allan Scientific) following the manufacturer's instructions.

**Gene cloning and rapid amplification of cDNA ends PCR (RACE–PCR).** RNA extraction from *Sepia* embryos at stages 24–26 and from *Limulus* stages 19–20 was done using TRIzol reagent (Ambion) following the manufacturer's instructions. cDNA synthesis was performed by an AMV reverse transcriptase (New England Biolabs) following the manufacturer's instructions. PCR amplification was carried using the following primers SepiaSoxEr, TGCTACCATGTTAGAAGTCATGCCT; SepiaSoxEf, GATTACCCTGATTACAAATACCAGCCC; SepiaSoxDf, CCACTACCAGCTCATAGCAACCATCAG; SepiaSoxDr, GGGCTTTGAGG GGTCAGGTTTCTCT; SepiaColAaf, AACGCCCCTGCCCGTTCCTGTCGC GATC; SepiaColAar, TCCCAATTCTATATGGAAGTCTTGT; Sepiahhf, TAATG TATCGGAAAACACAGTTGGTGCCA; Sepiahhr, GAGGAAGGCGATGA CTTCGCTGTAA; SepiabetacateninF1, TGTGCTGCTGGCATTCTGTCCAATC; SepiabetacateninR1, GCGACTCCTTCGTTCCTGGAGTGTA; LimulusSoxDF, CCAAAGAGAACTTGTATTGTGGATGGC; LimulusSoxDR, GGTGTCTG TCTCTCAGCTTGAAACATACCA; LimulusSoxEF, TTGCATGGACAA ACTCGTCAACTCGGT; LimulusSoxER, GGAAACTGGATACTGATGATAT GGAGTATC; LimulusColAaF, ATATGATGCAAGTGCTCTTGCTGCTCTCCT; LimulusColAaR, CTCACTGAAGAGTTGTAGGAAACTAAGCTG; LimulusHhF, GTCTTTAAGCARCAYGTNCCNAA; LimulusHhR, AAAGTTTGCGTACCART GDATNCC; LimulusbcatF, TTATGCCATCACTACCTTGCACAATCTC; LimulusbcatR, CTTGACAAGTGCAGGAATTCCCCCAGAT.

Full-length cDNA clones were isolated by rapid amplification of cDNA ends using SMARTer RACE 5′/3′ Kit (Clontech) and synthesis of 5′ and 3′ RACE cDNA libraries was performed following the manufacturer's instructions. The primers used in RACE–PCR experiments were as follows: HSCraceColAbr, GTAAAACGACGGCCA GTCGGCAGTGGTAGGTAATATTCTGTACAGC; SepPhRACEcolAbr, GTAAAACGACGGCCAGTCGGCAGTGGTAGGTAATATTCTGTACAGC; SepPhRACEcolAar, GTAAAACGACGGCCAGTGAGACAACCACACATAGGACTC TCCGGCT.

Sequences for *Sepia ColAa*, *ColAb*, *SoxE*, *SoxD*, *Hh* and *β-catenin*, and for *Limulus ColA*, *SoxE*, *SoxD*, *Hh* and *β-catenin*, have been deposited in GenBank under accession numbers KP322116–KP322126.

**ISH and immunohistochemistry.** Whole-mount ISH was performed using digoxigenin- and fluorescein-labelled antisense (or sense control) RNA probes according to protocols previously described for *Sepia*[27] and *Limulus*[33]. ISH on cryosections was performed using previously described protocols for vertebrate tissues[37]. PCNA, β-catenin and hyaluronan detection was performed on cryosections using mouse anti-PCNA (ab29, abcam), anti-β-catenin (C2206, Sigma-Aldrich) and biotinylated hyaluronic acid binding protein (385911, EMD Millipore). Hyaluronan detection on mouse, *Sepia* and *Limulus* cartilages was performed using streptavidin-HRP with Alexa Fluor 488 tyramide signal amplification (Molecular Probes).

**ISH and phalloidin staining.** Phalloidin staining was performed on cryosectioned embryos after whole-mount ISH. RNA expression was imaged by detecting the fluorescence generated by the NBT/BCIP precipitate emission (over 700 nm) when excited at 633 nm. Phalloidin staining was done using Alexa Fluor 488 phalloidin (Life Technologies). Sections were blocked with 1% BSA (A9647, Sigma-Aldrich) in PBS before a 30 min incubation with Alexa Fluor 488 phalloidin at 6.6 μM in blocking solution (1% BSA in PBS).

**Optical projection tomography of embryos after ISH.** *Sepia* and *Limulus* embryos were fixed in 4% paraformaldehyde in PBS after the completion of whole-mount ISH, and embryos were prepared and scanned following previously described protocols for vertebrate embryos[38,39]. Optical projection tomography scanning was performed using a Bioptonics 3001 OPT Scanner. The anatomy and gene expression channels were reconstructed using NRecon software and imported into the Amira program for three-dimensional visualization, analysis and renderings of three-dimensional images and videos.

**Molecular phylogenetic analysis of *collagen* and *Sox* genes.** Phylogenetic analyses of *collagen* and *Sox* sequences cloned from *Sepia* and *Limulus* cDNA pools were aligned with putative orthologues derived from EST databases (NCBI) by tBlastn searches using mouse *Col2a1* (AAH51383) and *Haliotis* collagen pro-alpha chain (BAA75668) collagens, and mouse *Sox9* (NP_035578) and *Sox6* (AAC52263). The retrieved sequences used for the phylogenetic analyses can be found in Supplementary Tables 1 and 2. Amino-acid sequences were aligned using MUSCLE[40] and phylogenetic reconstruction was performed with MrBayes 3.2.2 (ref. 41) using the WAG model[42] of amino-acid substitution, as described previously[43].

**Treatments with small-molecule inhibitors.** *S. officinalis* embryos were staged inside their egg cases after removing the outer layers of the egg case until the remaining inner layers were translucent enough to see the embryo. Embryos were selected for treatments once they reached stages 23–24, when they have already developed the funnel epithelium and associated mesenchyme (Extended Data Figs 7 and 8). Treatments were done in 100 ml glass beakers with 50 ml of sterile artificial seawater. Control embryos were treated with DMSO at 0.1%, and experimental embryos were exposed to 10 μM of cyclopamine (C988400, Toronto Research Chemicals), SANT-1 (S4572, Sigma-Aldrich), alsterpaullone (A4847, Sigma-Aldrich), BIO (B1686, Sigma-Aldrich), IWR-1 (I0161, Sigma), or PNU-74654 (P0052, Sigma-Aldrich) by adding 50 μl of 10 mM stock solutions for each of the drugs. DMSO was used as the solvent for all stock solutions. The following small-molecule inhibitors were used to target Hh and β-catenin signalling: cyclopamine and SANT-1 function as Smoothened inhibitors and antagonize Hh signalling; alsterpaullone and BIO function as agonists of β-catenin signalling by inhibiting GSK-3β (alsterpaullone has a broader spectrum and also inhibits other kinases, such as CDK1 and CDK5); IWR-1 and PNU-74654 work as antagonists of β-catenin signalling by inducing axin stabilization (stabilization of *β-catenin* destruction complex) and by blocking the interaction of β-catenin with Tcf, respectively.

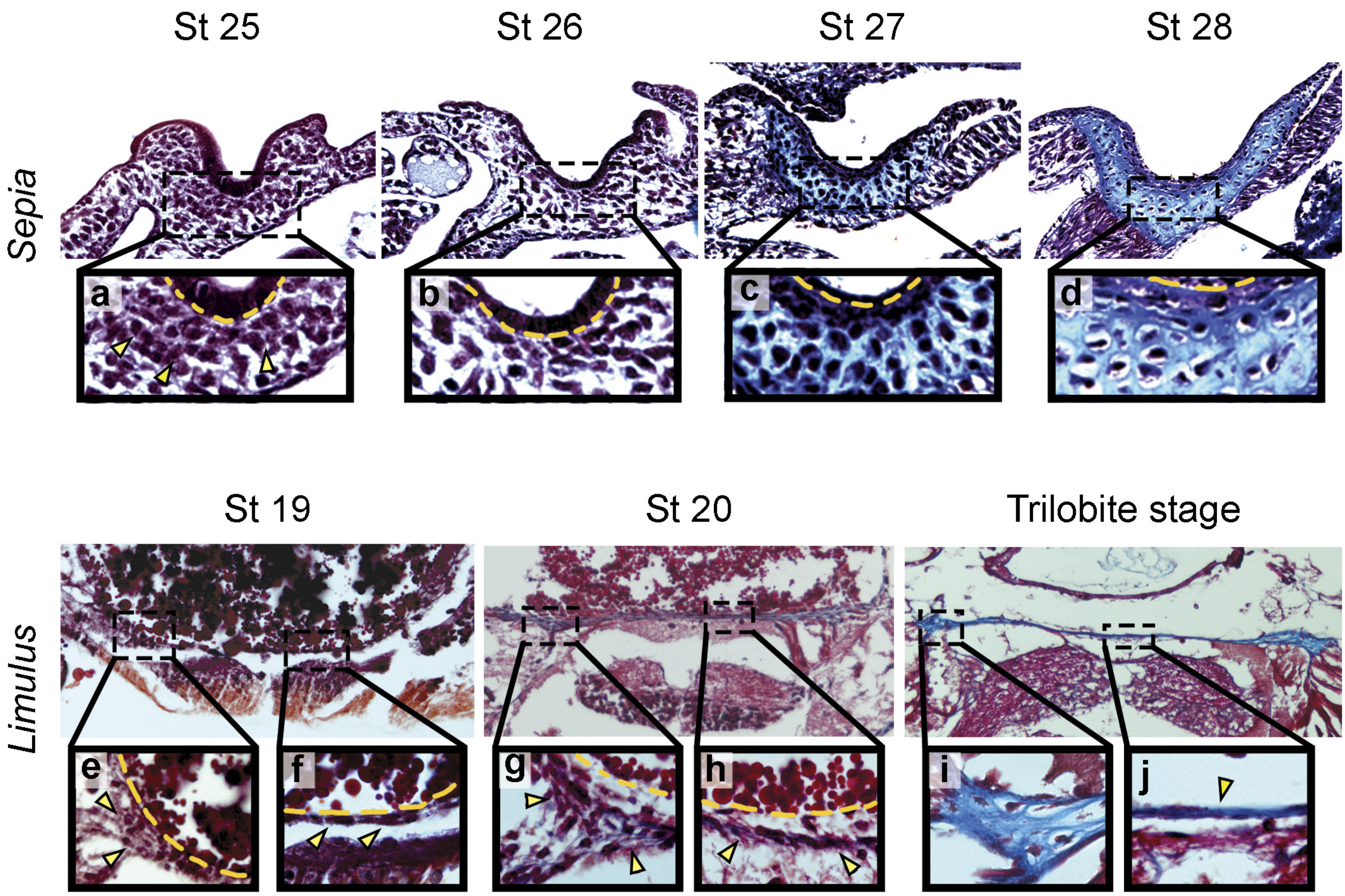
Embryos were incubated in 22 °C seawater with the treatment drug or with DMSO. Seawater containing the drug at the appropriate final concentration (or DMSO for controls) was replaced every 2 days, for a total exposure period of either 5 days or 10 days. Specimens were then collected and fixed for histology or ISH, as described above.

**BrdU labelling and BrdU pulse chase.** *S. officinalis* eggs ($n = 8$) were incubated in 0.05% BrdU in seawater at 22 °C for 6 h. Four embryos were processed immediately for immunofluorescence after BrdU incubation; the rest of the eggs were rinsed several times in seawater free of BrdU, then washed five times every 10 min, then five times every 30 min, and finally incubated for 3 days with complete water changes every 24 h. Embryos were fixed overnight in 4% paraformaldehyde. BrdU labelling was detected with an anti-BrdU antibody (G3G4, DSHB) and immunofluorescence on cryosections. Antigen retrieval was performed with incubation for 30 min in 2 N HCl.

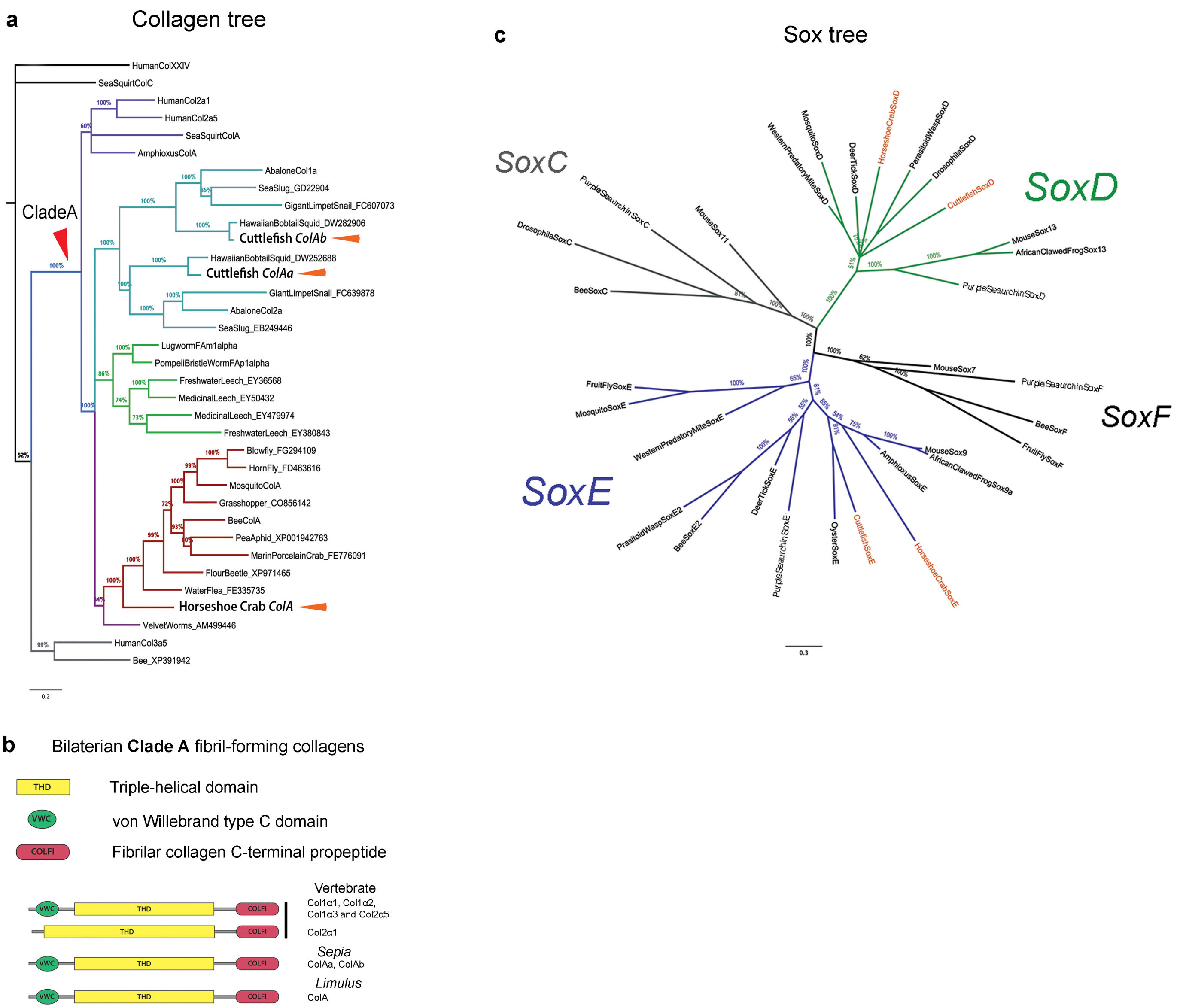
**Luciferase assay.** We used a firefly luciferase reporter construct controlled by the *Col2a1* promoter and the *Col2a1* chondrocyte-specific enhancer[44] (Extended Data Fig. 5f). We cloned full-length *SoxE* and *Sox9* from cDNA by reverse transcription PCR and ligated each into a pcDNA3.3 expression vector under the control of a CMV promoter (Invitrogen). Two cell lines, NIH3T3 mouse fibroblast (ATCC, CRL-1658) and IRC cells (a gift from W. Horton), were transfected using Lipofectamine 3000 (Invitrogen) with 200 ng of DNA per well (48-well plates) corresponding to the luciferase plasmid and the corresponding expression vector at a ratio of 1:3. We used the firefly luciferase vector and a Renilla luciferase control vector at a ratio of 20:1. After transfection, cells were cultured for 48 h as previously described[44] and luciferase activity was measured using a Dual-Luciferase Reporter Assay System (E1910, Promega) according to the manufacturer's instructions.

31. Lemaire, J. Table de developpement embryonnaire de *Sepia officinalis*. L. (mollusque cephalopode). *Bull. Soc. Zool. Fr.* **95,** 773–782 (1970).
32. Sekiguchi, K., Yamamichi, Y. & Costlow, J. D. Horseshoe crab developmental studies I. Normal embryonic development of *Limulus polyphemus* compared with *Tachypleus tridentatus*. *Prog. Clin. Biol. Res.* **81,** 53–73 (1982).
33. Blackburn, D. C. *et al.* Isolation and expression of Pax6 and atonal homologues in the American horseshoe crab, *Limulus polyphemus*. *Dev. Dynam.* **237,** 2209–2219 (2008).
34. Hall, B. K. *Bones and Cartilage: Developmental and Evolutionary Skeletal Biology* Ch. 4, 51–63 (Elsevier, 2005).

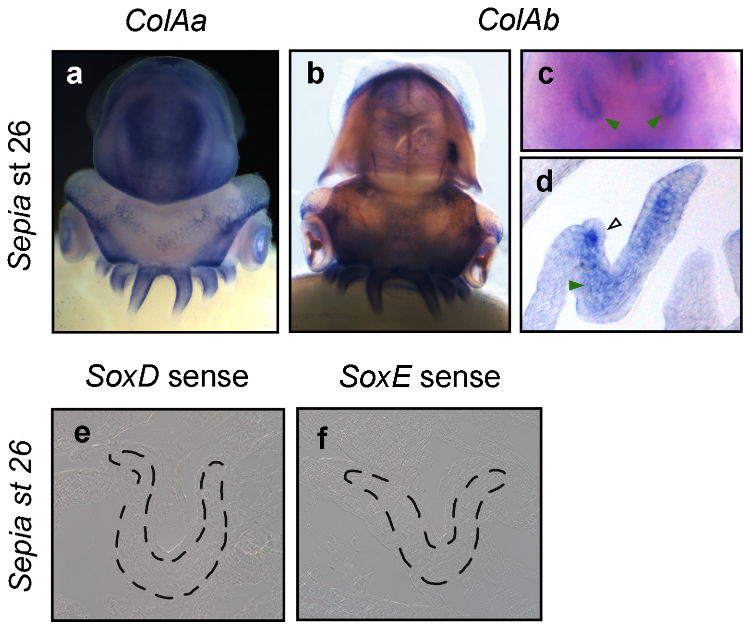
35. Sugahara, K. *et al.* Novel sulfated oligosaccharides containing 3-*O*-sulfated glucuronic acid from king crab cartilage chondroitin sulfate K. Unexpected degradation by chondroitinase ABC. *J. Biol. Chem.* **271,** 26745–26754 (1996).
36. Kinoshita, A. *et al.* Novel tetrasaccharides isolated from squid cartilage chondroitin sulfate E contain unusual sulfated disaccharide units GlcA(3-*O*-sulfate)β1–3GalNAc(6-*O*-sulfate) or GlcA(3-*O*-sulfate)β1–3GalNAc. *J. Biol. Chem.* **272,** 19656–19665 (1997).
37. Abzhanov, A. Darwin's finches: analysis of beak morphological changes during evolution. *Cold Spring Harbor Protoc.* **2009,** emo119 (2009).
38. Quintana, L. & Sharpe, J. Preparation of mouse embryos for optical projection tomography imaging. *Cold Spring Harb. Protoc.* **2011,** 664–669 (2011).
39. Quintana, L. & Sharpe, J. Optical projection tomography of vertebrate embryo development. *Cold Spring Harb. Protoc.* **2011,** 586–594 (2011).
40. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32,** 1792–1797 (2004).
41. Huelsenbeck, J. P. & Ronquist, F. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17,** 754–755 (2001).
42. Whelan, S. & Goldman, N. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol.* **18,** 691–699 (2001).
43. Zhang, G., Miyamoto, M. M. & Cohn, M. J. Lamprey type II collagen and Sox9 reveal an ancient origin of the vertebrate collagenous skeleton. *Proc. Natl Acad. Sci. USA* **103,** 3180–3185 (2006).
44. Lefebvre, V., Huang, W., Harley, V. R., Goodfellow, P. N. & de Crombrugghe, B. SOX9 is a potent activator of the chondrocyte-specific enhancer of the pro alpha1(II) collagen gene. *Mol. Cell. Biol.* **17,** 2336–2346 (1997).



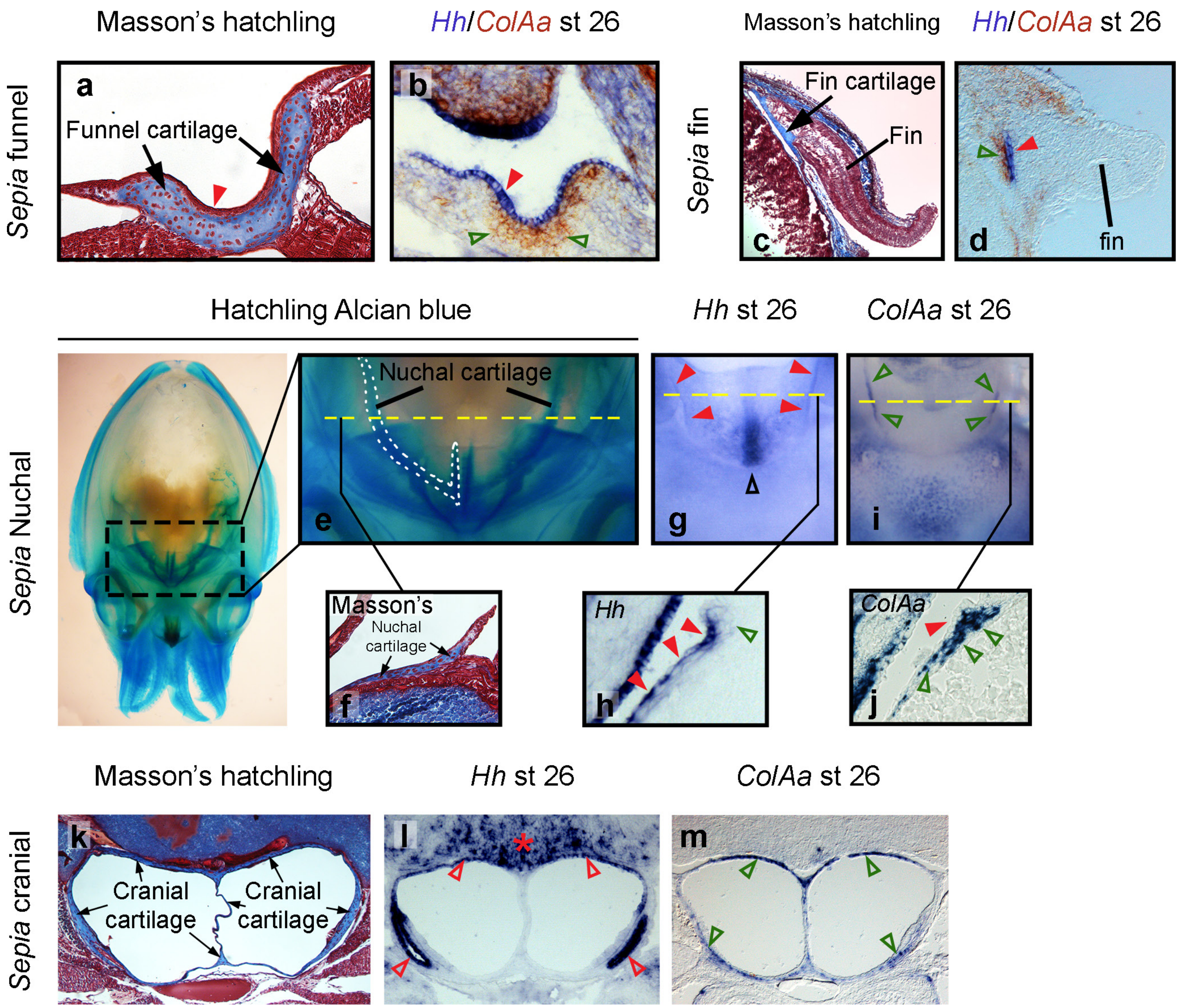

**Extended Data Figure 1 | Developmental series showing chondrogenesis in *Sepia* and *Limulus*.** Masson's trichrome-stained sections. Collagen is stained blue. **a**–**d**, Sections through funnel cartilage of *Sepia* embryos. Bottom row shows high magnification of boxed area. Yellow arrowheads mark the pre-cartilaginous cell condensation and the yellow dashed line marks the level of the basal lamina of the funnel epithelium. **e**–**j**, Transverse sections through the endosternite of *Limulus* embryos. Bottom row shows high magnification of boxed area. Yellow arrowheads mark the pre-cartilaginous cell condensations and the yellow dashed line delineates the mesenchyme from the yolk cavity.



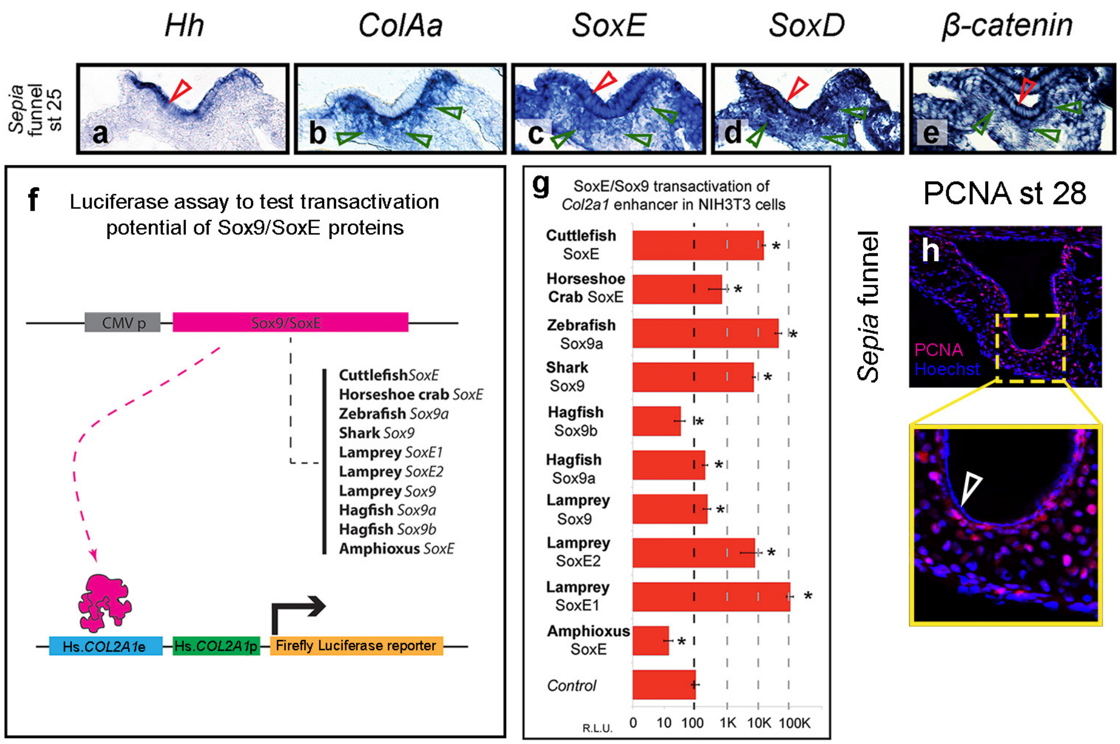

**Extended Data Figure 2 | Molecular phylogenetic analysis of clade A fibrillar collagens and Sox transcription factors (*SoxC, SoxD, SoxE* and *SoxF*). a**, Molecular phylogeny clade A fibrillar collagens (*ColA*) using the carboxy (C)-terminal propeptide shows that that *ColA* genes are represented in all major lineages of Bilateria (Deuterostomia, purple; Annelida, green; Mollusca, cyan; Arthropoda, red) and indicates that *Sepia* and *Limulus* (orange arrowheads) sequences belong to the *ColA* family (see Supplementary Table 1 for sequence accession numbers). **b**, Shared architecture of ColA propeptide between vertebrates and protostome invertebrates. In vertebrates, the von Willebrand type C domain is absent in Col2α1 but present in the other clade A collagens (Col1α1, Col1α2, Col1α3 and Col2α5). **c**, Molecular phylogeny of *Sox* genes using the HMG DNA binding domain under the WAG amino-acid model of evolution. The sequences derived from *Sepia* and *Limulus* (in orange) belong to the SoxE and SoxD families (see Supplementary Table 2 for sequence access numbers). All trees were generated by Bayesian phylogenetic inference using WAG model of amino-acid substitution. Branch support shown as percentage of posterior probabilities.



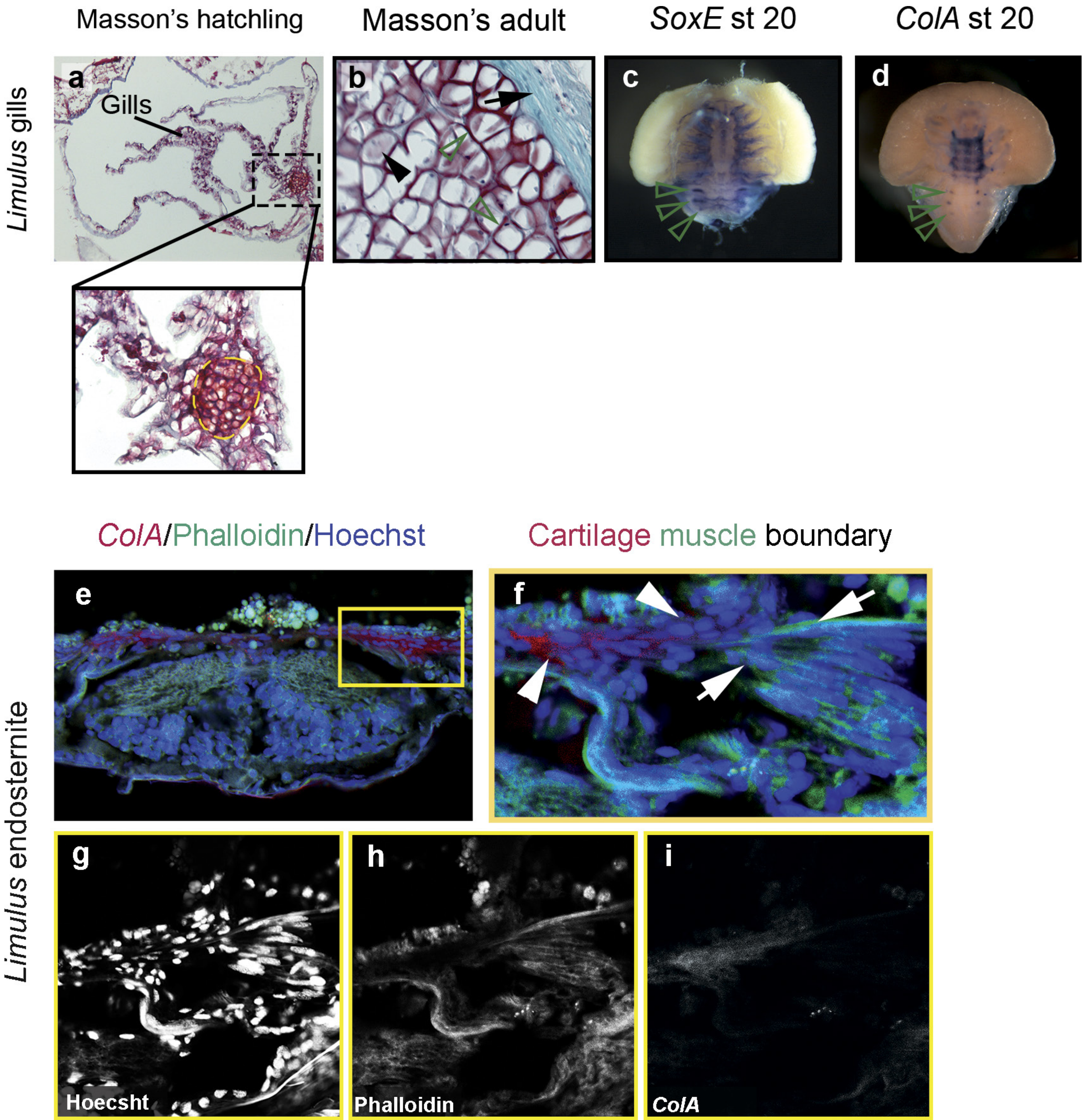

**Extended Data Figure 3 | *ColAa* and *ColAb* show similar patterns of gene expression in *Sepia* embryos. a**, **b**, Whole-mount ISH for (**a**) *ColAa* and (**b**) *ColAb*. Dorsal views. **c**, Ventral view of *ColAb* ISH showing the funnel cartilage precursors, marked by green arrowheads. **d**, Cryosections of these embryos reveal that *ColAb* is expressed in pre-chondrogenic mesenchyme (green arrowhead). Funnel epithelium is marked by black open arrowhead. **e**, **f**, Negative control ISH for (**e**) *SoxD* and (**d**) *SoxE* using sense RNA probes; broken lines outline pre-chondrogenic cells that form the funnel cartilage.



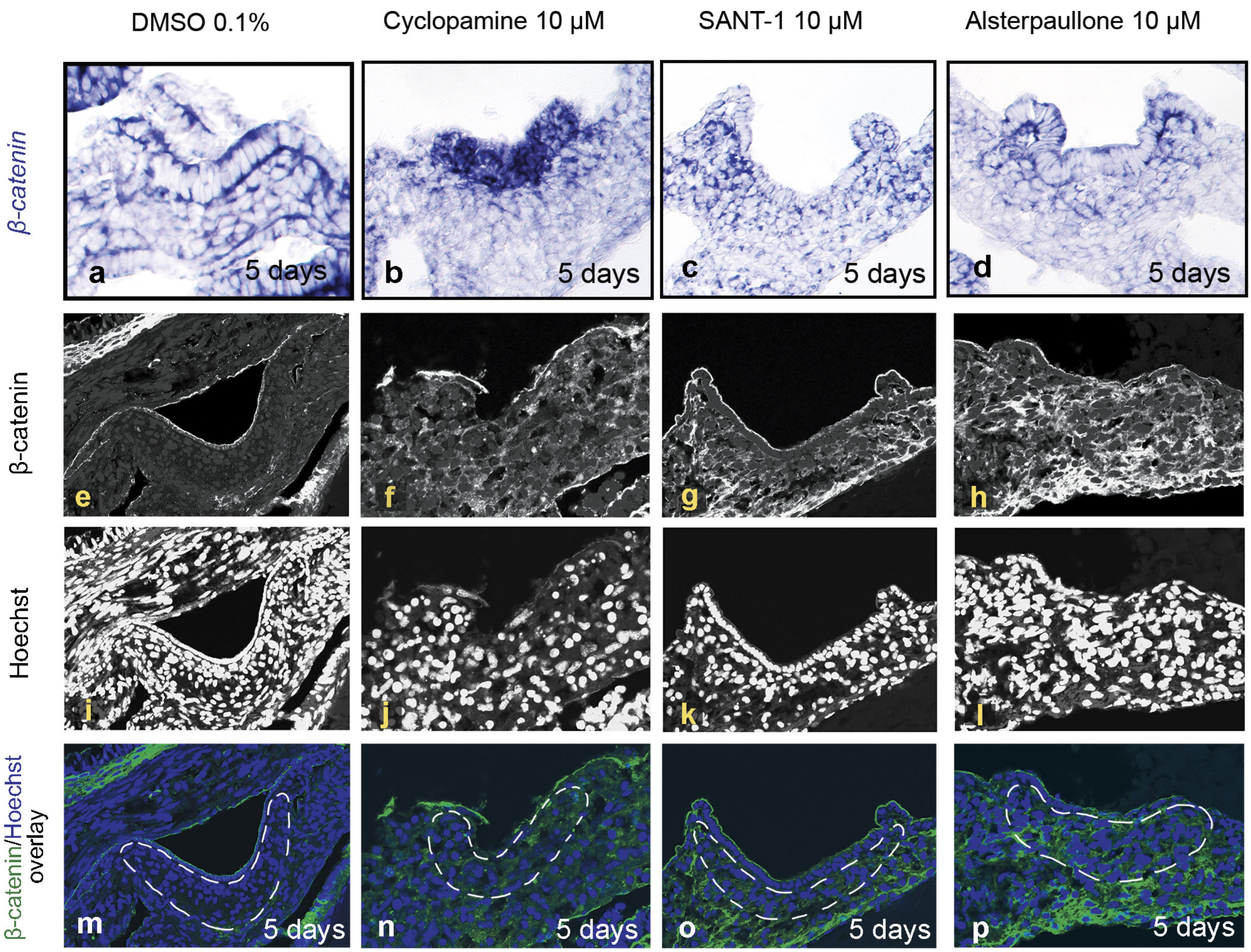

**Extended Data Figure 4 | Chondrogenesis of multiple cartilages occurs near *Hedgehog*-expressing tissues in *Sepia*. a**, Funnel cartilage in a hatchling of *Sepia* (black arrows) located underneath the funnel epithelium (red arrowhead). **b**, Double ISH of the funnel cartilage primordium at stage 26, showing the expression of *ColAa* (brown stain) in pre-cartilaginous cells (green arrowheads) and *Hedgehog (Hh*; purple stain) in the funnel epithelium (red arrowhead). **c**, Fin cartilage located at the base of the fin (black arrows) in a hatchling. **d**, Double ISH of the fin at stage 26 showing pre-cartilaginous mesenchyme expressing *ColAa* (brown stain, green arrowheads) next to a *Hh* domain (purple stain, red arrowhead). **e**, Whole-mount alcian-blue-stained *Sepia* hatchling. The white dashed outline marks the right nuchal cartilage and the yellow dashed line indicates the approximate plane of the section shown in **f**, which is stained with Masson's trichrome. **g**, Whole-mount ISH showing *Hh* expression on the right and left nuchal cartilage primordia at stage 26 (red arrowheads). A large domain of *Hh* expression can also be observed in the midline (black open arrowhead) between the nuchal cartilage primordium. Yellow dashed line in **g** indicates approximate plane of section shown in **h**, a cryosection showing the expression of *Hh* in the epithelium of the nuchal cartilage primordium (red arrowheads) but not in the mesenchyme (green open arrowhead). **i**, Whole-mount ISH of *ColAa* at stage 26 showing its expression on the nuchal cartilage primordia (green open arrowheads). Yellow dashed line in **i** indicates approximate plane of section shown in **j**, which shows a cryosection showing the expression of *ColAa* in the mesenchyme (green open arrowheads) of the nuchal cartilage primordium, but not in the epithelium (red arrowhead). **k**, Histological section stained with Masson's trichrome at the level of the paired statocyst cavities surrounded by cranial cartilages. **l**, ISH on cryosections from a stage 26 embryo reveal that the brain (marked by a red asterisk) and most of the inner epithelial lining of the statocyst cavities express *Hh* (red open arrowheads). **m**, The pre-cartilaginous cells underneath the *Hh* domain express *ColAa* (marked by green open arrowheads).



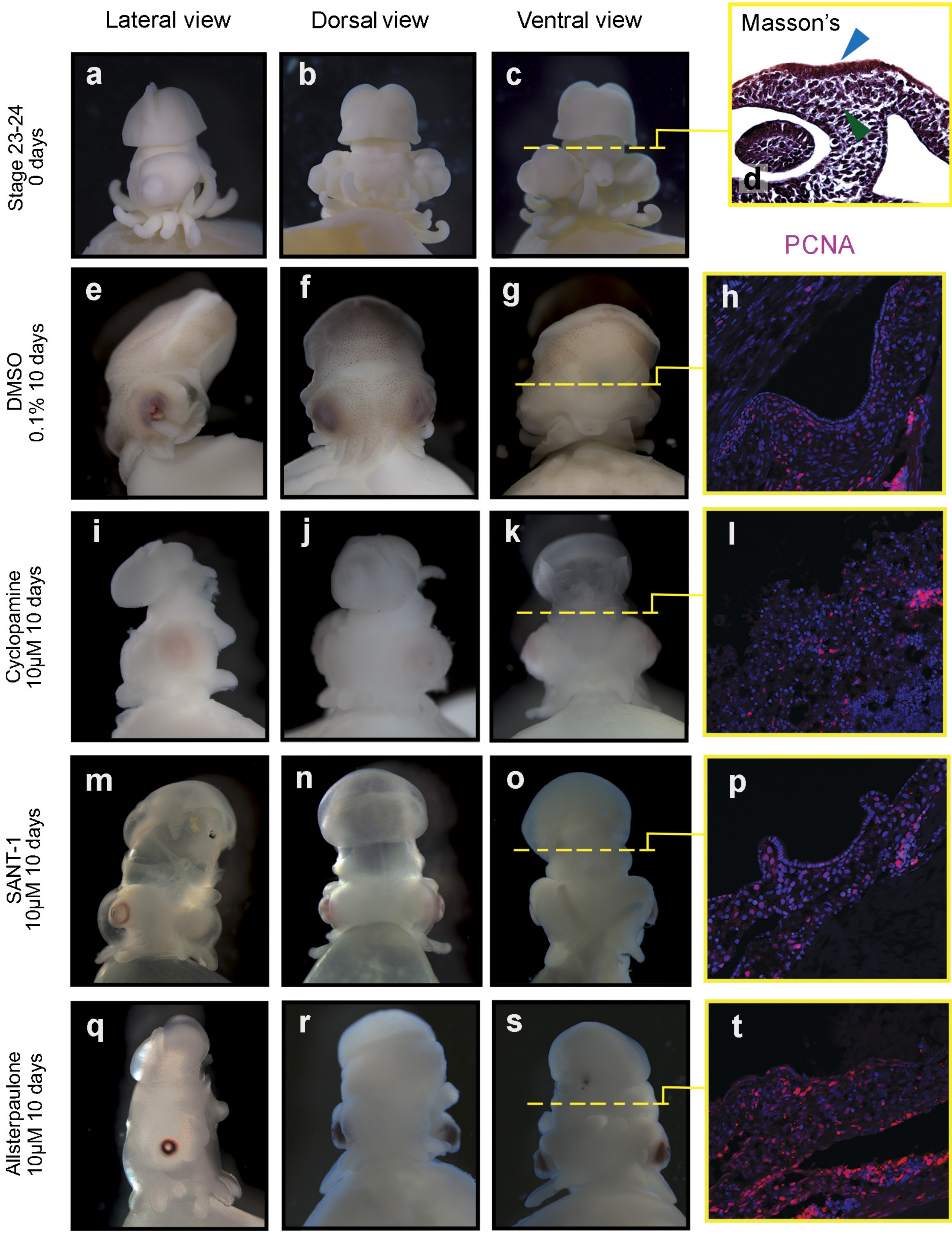

**Extended Data Figure 5 | Patterns of gene expression in developing funnel cartilage of *Sepia* at stage 25. a**, *Hh* is expressed in the funnel epithelium. **b**, *ColAa* is expressed in pre-cartilaginous cells. **c**, *SoxE* is expressed in the funnel epithelium as well as in the pre-cartilaginous cells, similar to **d**. **d**, **e**, *SoxD* (**d**) and *β-catenin* (**e**) expression in the funnel cartilage progenitors. In all figures, red open arrowheads mark the funnel epithelium and green open arrowheads mark pre-cartilaginous cells. **f**, Schematic representation of the luciferase reporter assay to test the transactivation potential of Sox9/SoxE transcription factors. Cells were co-transfected with a Sox9/SoxE expression vector under the control of a ubiquitous CMV promoter. The luciferase reporter was controlled by upstream *Col2a1* regulatory elements, four tandem copies of the chondrocyte-specific human *Co2a1* enhancer, and the human *Col2a1* promoter. **g**, SoxE and Sox9 transactivation of the human *Col2a1* enhancer in NIH3T3 mouse fibroblast cells, assayed by the activity of a luciferase reporter driven by the *Col2a1* enhancer. Asterisks indicate significant differences over control levels (*t*-test; $P \leq 0.05$); error bars, s.d. Each luciferase experiment was repeated four times, with four replicates per experiment. **h**, PCNA immunofluorescence in the mature funnel cartilage of stage 28 embryos indicates active proliferation in the chondrocytes over the entire cartilaginous element (bottom panel shows high magnification of boxed area above; white open arrowhead marks the epithelium). Proliferation becomes restricted to the sub-epithelial layer one stage later (compare with Fig. 3).



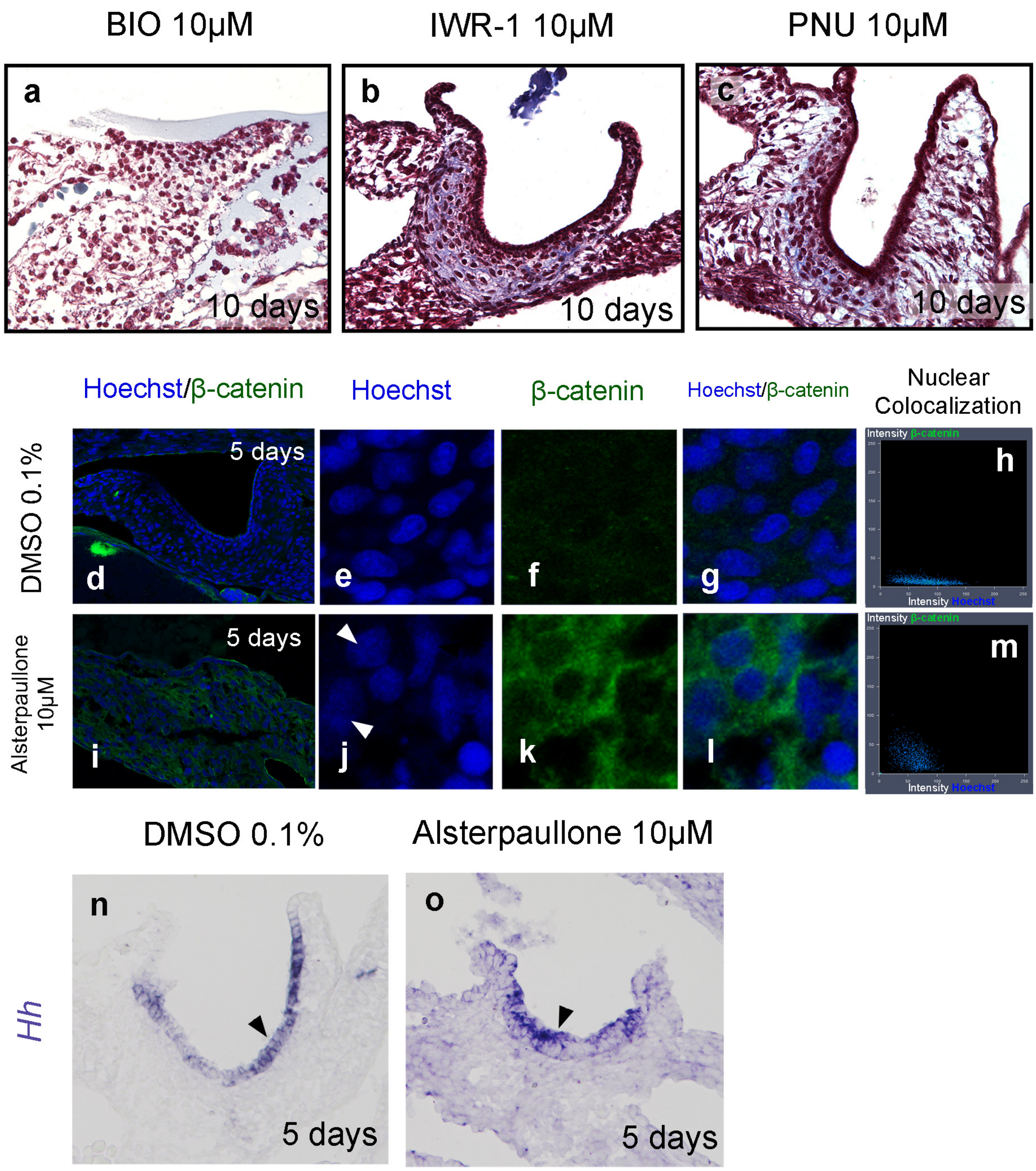

**Extended Data Figure 6 | Gill and endosternite cartilages in *Limulus* are collagen-based and express *SoxE* during chondrogenesis. a**, Section through gills of *Limulus* hatchlings stained with Masson's trichrome. Gill cartilage is located at the base of the gills (outlined by yellow dashed lines). **b**, Adult gill cartilage stained with Masson's trichrome showing a cell-rich tissue with hypertrophic cells (black arrowhead) separated by thin extracellular matrix (black open arrowheads); the gill cartilage ECM shows no aniline blue stain compared with the surrounding connective tissue; however, during embryonic development *SoxE* (**c**) and *ColA* (**d**) are expressed in the gill cartilage primordia (green open arrowheads). **e**–**i**, Confocal imaging of endosternite after phalloidin staining and *ColA* ISH. **f**, Higher magnification of the boxed area in **e** showing the boundary between *ColA*-expressing pre-chondrogenic cells (white arrowheads) and the differentiating muscle cells (white arrows) attached to the endosternite pre-chondrogenic tissue. **g**–**i**, Separate channels from **f**, showing Hoescht (**g**), phalloidin (**h**) and *ColA* (**i**).




**Extended Data Figure 7 | Expression of β-catenin transcripts and protein after 5-day treatments with cyclopamine, SANT-1, alsterpaullone and DMSO (control). a**–**d**, After treatment for 5 days with small-molecule inhibitors, *β-catenin* transcripts can be detected in the funnel cartilage primordium in the (**a**) DMSO controls as well as in (**b**) cyclopamine-, (**c**) SANT-1- and (**d**) alsterpaulone-treated embryos. **e**–**p**, In contrast to *β-catenin* mRNA, β-catenin protein is degraded during normal funnel chondrogenesis, as seen in the DMSO control (**e**, **m**); however, β-catenin protein remains in funnel chondroprogenitors after treatment with cyclopamine (**f**, **n**), SANT-1 (**g**, **o**) and alsterpaullone (**h**, **p**).




**Extended Data Figure 8 | Bright-field micrographs and immunofluorescence of *Sepia* embryos before and after treatments with the small-molecule inhibitors cyclopamine, SANT-1 and alsterpaullone, or with DMSO vehicle control. a–c**, *Sepia* embryos at the beginning of drug treatments (stages 23–24). **d**, Histological sections at the beginning of the treatments demonstrating the presence of the funnel epithelium and the associated mesenchyme. The cuboidal signalling epithelium (blue arrowhead) and pre-cartilaginous mesenchyme (green arrowhead) can be identified. **e–t**, *Sepia* embryos after 10 days of treatment: **e–g**, control DMSO-treated embryos; **i–k**, cyclopamine-treated embryos; **m–o**, SANT-1-treated embryos; **q–s**, alsterpaullone-treated embryos. **h**, **l**, **p**, **t**, PCNA staining shows that cell proliferation in funnel cartilage continued after drug treatments, indicating that treatments did not induce global toxicity. DMSO control (**h**) cyclopamine- (**l**) and SANT-1-treated (**p**) embryos stained positive for cell proliferation in the funnel cartilage, and alsterpaullone-treated (**t**) embryos showed stronger PCNA staining of the funnel cartilage than did cyclopamine-treated embryos, SANT-1-treated embryos or DMSO controls.




**Extended Data Figure 9 | Positive and negative modulation of β-catenin signalling has opposite effects on chondrogenesis in *Sepia*.** **a**, Stabilization of β-catenin signalling using the GSK-3β inhibitor BIO prevents funnel cartilage development, as revealed by Masson's trichrome. **b**, **c**, Inhibition of β-catenin signalling by inducing axin stabilization (stabilization of β-catenin destruction complex) with IWR-1 (**b**) or by blocking the interaction of β-catenin and Tcf with PNU (**c**), did not disrupt chondrogenesis of funnel cartilage. **d**–**g**, **i**–**l**, Cellular accumulation of β-catenin in funnel cartilage of alsterpaullone-treated embryos compared with DMSO controls. β-catenin nuclear localization is not observed in DMSO control embryos (**d**–**g**), but after alsterpaullone treatment, β-catenin accumulates in the cytoplasm and the nucleus (**i**–**l**). Arrowheads mark two cells stained with Hoechst (**j**) that are rich in β-catenin (**k**). **g**, **l**, Overlay of Hoechst/β-catenin from **e** and **f** (**g**) and **j** and **k** (**l**). **h**, **m**, Nuclear co-localization plots of funnel cartilage cells showing β-catenin intensities in Hoechst-positive domains (nuclei); cytoplasmic β-catenin signal is not plotted. Alsterpaullone-treated embryos (**m**) show higher β-catenin intensities than DMSO controls (**h**), demonstrating β-catenin accumulation in the nuclei. **n**–**o**, Accumulation of β-catenin does not affect Hh expression in the funnel epithelium after alsterpaullone treatments; compare **o** with DMSO controls in **n**.

LETTER

doi:10.1038/nature17941

# Topology of ON and OFF inputs in visual cortex enables an invariant columnar architecture

Kuo-Sheng Lee[1,2], Xiaoying Huang[1] & David Fitzpatrick[1]

**Circuits in the visual cortex integrate the information derived from separate ON (light-responsive) and OFF (dark-responsive) pathways to construct orderly columnar representations of stimulus orientation and visual space[1–7]. How this transformation is achieved to meet the specific topographic constraints of each representation remains unclear. Here we report several novel features of ON–OFF convergence visualized by mapping the receptive fields of layer 2/3 neurons in the tree shrew (*Tupaia belangeri*) visual cortex using two-photon imaging of GCaMP6 calcium signals. We show that the spatially separate ON and OFF subfields of simple cells in layer 2/3 exhibit topologically distinct relationships with the maps of visual space and orientation preference. The centres of OFF subfields for neurons in a given region of cortex are confined to a compact region of visual space and display a smooth visuotopic progression. By contrast, the centres of the ON subfields are distributed over a wider region of visual space, display substantial visuotopic scatter, and have an orientation-specific displacement consistent with orientation preference map structure. As a result, cortical columns exhibit an invariant aggregate receptive field structure: an OFF-dominated central region flanked by ON-dominated subfields. This distinct arrangement of ON and OFF inputs enables continuity in the mapping of both orientation and visual space and the generation of a columnar map of absolute spatial phase.**

Circuits in the visual cortex transform the inputs supplied by ON- and OFF-centre axons from the lateral geniculate nucleus into a columnar architecture that preserves the orderly mapping of visual space, while generating *de novo* an iterated map of stimulus orientation[1–5]. The first step in this process involves the convergence of ON and OFF inputs onto single cortical neurons to create 'simple' receptive fields that exhibit spatially offset ON and OFF subfields[1,6,7]. Understanding the logic that cortical circuits use to integrate the ON and OFF pathways in order to build this columnar architecture requires the ability to visualize the receptive fields of large numbers of simple cells, determine the spatial arrangement of their ON and OFF subfields, and understand how this arrangement relates to the columnar maps of orientation and visual space. We have achieved this by using two-photon calcium imaging to map the receptive fields of large numbers of single neurons in layer 2/3 of the tree shrew, a species that has a close phylogenetic relation to primates[8] and a visual cortex with a well-developed columnar architecture[5,9–11].

Previous studies in the tree shrew have shown that ON- and OFF-centre inputs from the lateral geniculate nucleus target separate populations of neurons in cortical layer 4, and that the projections from layer 4 to layer 2/3 bring about the convergence of ON and OFF inputs onto single layer 2/3 neurons[12,13]. Here we mapped the receptive fields of single layer 2/3 neurons with reverse correlation to a sparse noise visual stimulus and analysed the spatial distribution of ON and OFF responses[14] (Extended Data Fig. 1; all statistical details can be found in Supplementary Notes). Layer 2/3 neurons exhibited robust responses to the sparse noise stimulus, making it possible to reliably reconstruct the receptive fields of hundreds of single layer 2/3 neurons per region of interest (generally 0.36–1.0 $mm^2$) (Fig. 1a, b), A large fraction (42%) of the layer 2/3 neurons that showed statistically significant receptive fields in response to this stimulus (see Methods) had spatially offset ON and OFF subfields, consistent with the organization of simple cells that has been described in other species. Other layer 2/3 neurons exhibited single-sign receptive fields (16% ON, 33% OFF) and a relatively small percentage appeared to have complex receptive fields with overlapping ON and OFF responses (9%) (Extended Data Fig. 2).

As a first step in understanding the transformation that underlies cortical columnar architecture, we examined the distribution in visual space of the receptive fields and the ON and OFF subfields of the simple cells that were found in a 1-$mm^2$ field of view. The receptive field centres of the neurons in the field of view were displaced over about 5° of visual space, an angle consistent with previous studies of the mapping of visual space in this species (average cortical magnification factor 0.2 mm per degree; refs 5, 10). However, the centres of the ON and OFF subfields exhibited a strikingly different distribution in visual space (Fig. 1c): the centres of the OFF subfields were clustered within a compact region of visual space, whereas the centres of the ON subfields were spread over a greater region of visual space, distributed around the region occupied by the OFF subfield centres. To quantify this difference, we computed the ratio of mean pairwise distances within each group, and compared this with the results found after shuffling polarity identity (Fig. 1d). These results revealed a fundamental difference in the visuotopic mapping of ON and OFF inputs in layer 2/3: the ON inputs that contribute to neural responses in a given region of the cortex originate from a broader region of visual space than do the OFF inputs. Epi-fluorescence imaging of population responses produced patterns of cortical activation that are consistent with these observations: a significantly greater area of the cortical surface is activated in response to stimulation by a light bar than by a dark bar of the same size (Extended Data Fig. 3).

Next we evaluated the precision of the visuotopic mapping of the receptive fields and the ON and OFF subfields of neurons with simple receptive fields in a given region of visual cortex. The receptive field centres of the neurons in a 1-$mm^2$ field of view always exhibited clear progressions in both azimuth and elevation. The centres of the OFF subfields also exhibited systematic progressions in both dimensions, and these were even more regular than those observed for receptive field centres. By contrast, the ON subfields from the same population exhibited a striking degree of disorder: adjacent neurons frequently had receptive fields with quite different ON subfield centre locations and there was little sign of fine visuotopic progression (Fig. 2a–c). Similar results were found for all eight sampled cortical regions for both simple (Fig. 2d, e) and single-sign cells (Extended Data Fig. 4). The visuotopic progression of OFF subfields is consistent with that predicted for a smooth visuotopic map with a deviation less than 1°, whereas the pattern of ON subfields cannot be explained by a smooth visuotopic progression. We conclude that the OFF inputs to layer 2/3

[1]Department of Functional Architecture and Development of Cerebral Cortex, Max Planck Florida Institute for Neuroscience, Jupiter, Florida 33458, USA. [2]Integrative Biology and Neuroscience Graduate Program, Florida Atlantic University, Boca Raton, Florida 33431, USA.

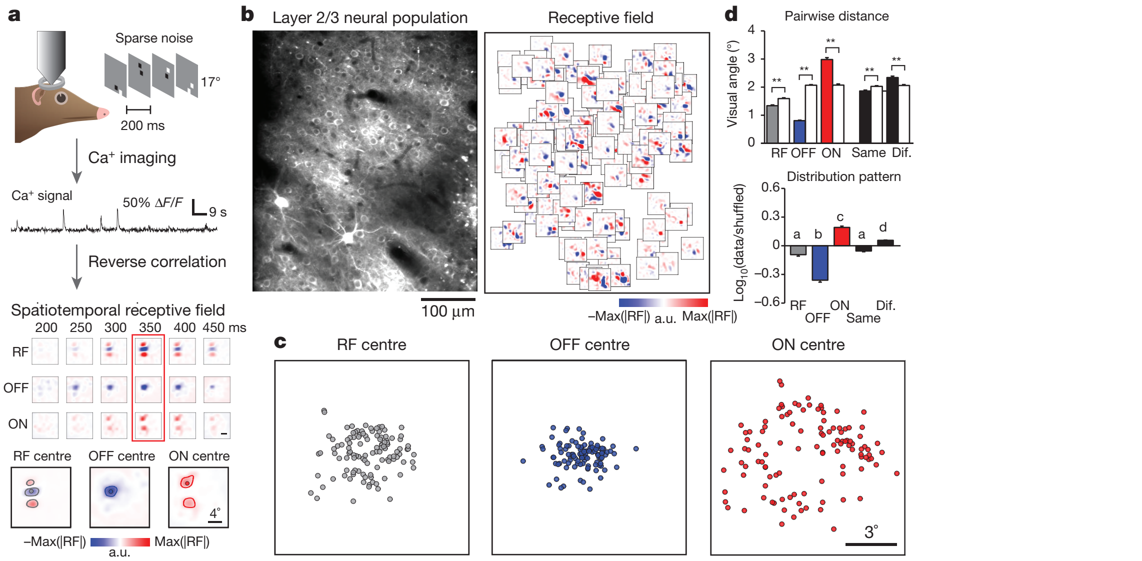



**Figure 1 | Differential arrangement of simple cell ON and OFF subfields in visual space.** **a**, Spatiotemporal receptive fields and ON or OFF subfields of cortical neurons were independently obtained using calcium imaging combined with reverse correlation of responses to a sparse noise stimulus. The receptive field and ON or OFF subfields were defined at the peak signal-to-noise ratio (SNR) time window. Small circles indicate the centres of mass of the whole receptive field and the ON and OFF subfields (see Methods). **b**, An example of a two-photon field of view (left) and all the significant receptive fields (same scale as receptive fields in **a**) from individual cells overlaid on their soma locations (right). **c**, An example of the distribution of receptive fields and OFF and ON subfield centres in visual space. **d**, The pairwise distance between the centres of mass for all categories and for shuffled data (white bars) from the example in **c**. Receptive fields, OFF subfields, and subfields sharing the same signs are more clustered, and ON subfields and subfields with different signs are more scattered, than by chance (rank-sum test for each group, $^{**}P < 0.0001$). The bottom plot summarizes the comparison of real and shuffled data; positive values indicate a scattered distribution pattern and negative values indicate a clustered distribution pattern relative to random shuffles ($n = 8$ fields of view from 7 animals; $P = 6.1 \times 10^{-21}$, Kruskal–Wallis test with post-hoc testing using Dunn's method; letters indicate groups with statistically significant difference of $P < 0.01$; see Methods). Error bars indicate s.e.m.

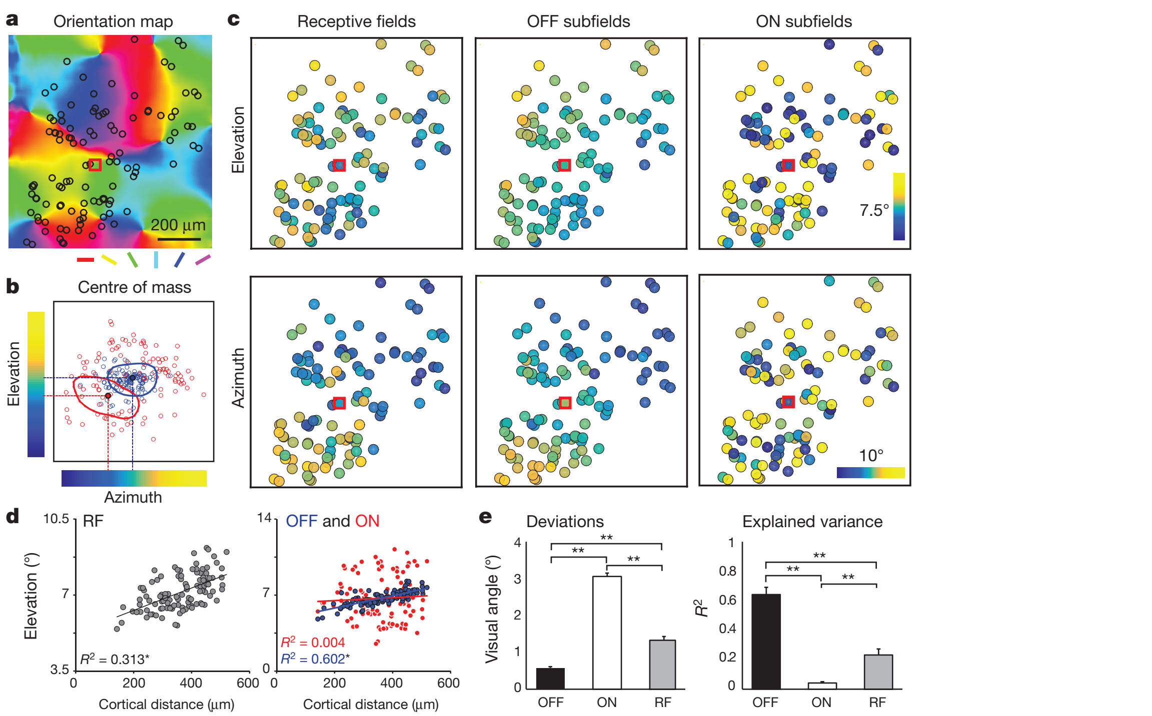



**Figure 2 | Differences in visuotopic precision of simple cell ON and OFF subfield centres.** **a**, Example field of view showing somatic location of all simple cells from four cortical depths (black circles) superimposed on an orientation map. Coloured bars at bottom show orientations. **b**, The locations in visual space of the centres of mass of ON and OFF subfields for neurons in **a**, illustrating the colour code used to depict azimuth and elevation values in **c**. The square represents the receptive field of an example cell (red square in **a** and **c**) showing translation of ON and OFF subfield centres into elevation–azimuth coordinates. **c**, The visual field locations (elevation and azimuth) for the receptive fields and ON and OFF subfields of each neuron illustrated in **a**. **d**, The relationship between cortical distance and position in visual space along the elevation axis for the receptive fields (left) and ON and OFF subfields (right) from the example in **c** (linear regression, $^{*}P < 0.0001$). **e**, Summary of the deviations of the experimental data from smooth visuotopy (left) and the degree to which smooth visuotopy accounts for the variance in the experimental data (right) (Kruskal–Wallis test with post-hoc testing using Dunn's method, $n = 16$ visuotopic maps combining elevation and azimuth results from 8 imaging areas, $^{**}P < 0.0001$; see Methods). Error bars indicate s.e.m.

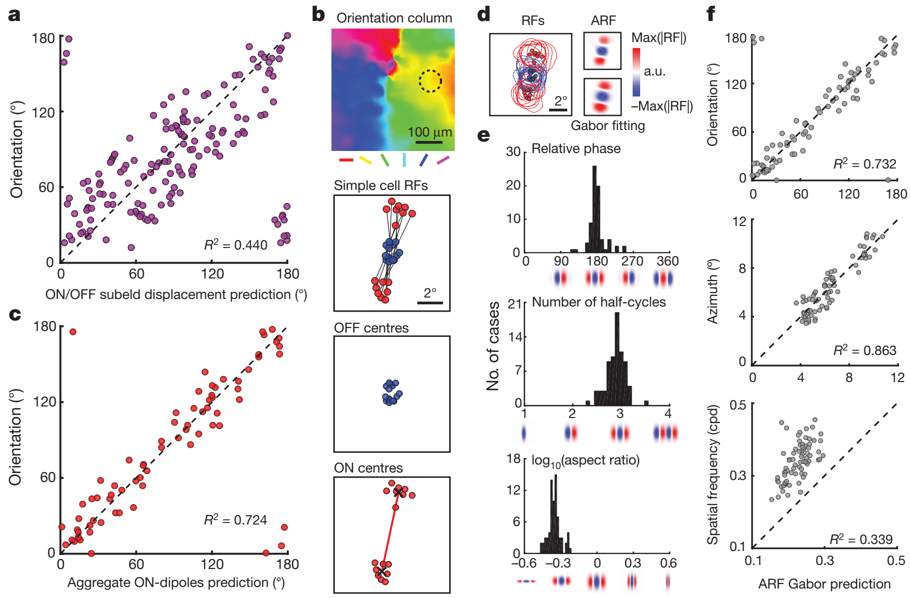



**Figure 3 | Orientation columns exhibit an invariant aggregate receptive field structure.** **a**, Consistent with simple cells in other mammals, the displacement of ON and OFF subfields in visual space predicts the preferred orientation in individual cells (linear regression, $n = 176$ cells from 2 animals, $P < 0.0001$). **b**, Example of receptive fields (RFs) from the simple cells in a single orientation column (dashed circle). Lines connect the ON subfield (red) and the OFF subfield (blue) centres of individual simple cell receptive fields. The ON centres form two clusters that define the aggregate ON-dipole of the column. **c**, The aggregate ON-dipoles from all the simple cells within individual orientation columns predict the preferred orientation of the column (linear regression, $P < 0.0001$). **d**, The normalized simple cell receptive fields from a single column in **b** were averaged to derive the aggregate receptive field (ARF) which was fit with a Gabor function. **e**, Cortical columns exhibit an invariant ARF structure resembling an OFF-centred simple cell receptive field with a specific relative phase, number of half-cycles, and aspect ratio. **f**, The parameters of the ARF Gabor fit account for multiple features of the cortical column including orientation, visual position, and spatial frequency ($n = 73$ cortical columns from 5 animals; circular or linear regression, all $P < 0.0001$; see Methods).

neurons are arranged with fine visuotopic precision that is absent for ON inputs.

In most simple cells in other species, the visuotopic displacement of the ON and OFF subfields is correlated with a cell's orientation preference: subfields are displaced along an axis in visual space that is orthogonal to the cell's preferred orientation[1,4,15–17]. This correlation suggests that the disorderly visuotopic arrangement of ON subfields in layer 2/3 neurons might be explained by their orderly arrangement in relation to the map of orientation preference. To test this possibility, we first compared the preferred orientation of individual neurons to that predicted by the axis of displacement of the ON and OFF centre subfields and verified this strong correlation at the level of single neurons (Fig. 3a). Next, we asked how well the visuotopic locations of the ON subfield centres of the neurons in a column (within a diameter of 80 μm, orientation difference within 11.25° of the mean, see Methods) predicted the preferred stimulus orientation of the column. The locations of the ON subfield centres for the simple cells in a column are highly clustered in visual space, forming dipoles that strongly predict the orientation preference of the column (Fig. 3b, c). Thus both the ON and OFF pathways exhibit a high degree of precision in their topological arrangement but for different columnar maps: the OFF pathway exhibits precision for the map of visual space while the ON pathway exhibits precision for the map of orientation preference. The distinct topological arrangement of both ON and OFF subfield centres that we have demonstrated for individual columns is maintained at the single neuron level, irrespective of location in the orientation map (Extended Data Fig. 5).

These observations suggest that all orientation columns have simple cells that are arranged with a fundamentally similar visuotopic structure: an OFF-dominated central region flanked by ON-dominated subfields. To test this possibility, we computed the aggregate receptive field (ARF) for a cortical column by simply overlaying the ON and OFF subfields for each neuron in the column on the basis of their positions in visual space (Extended Data Fig. 6). Individual neurons in a column exhibit different subfield organizations, with an average of 2–3 subfields. Overlaying all the simple cell receptive fields in a column results in a column aggregate receptive field that is somewhat larger than the receptive fields of the individual neurons in the column, but clearly exhibits an OFF-dominated central region flanked by two ON-dominated regions. In essence, the aggregate receptive field of a column resembles the receptive field of an OFF-centred simple cell and it can be well fit by a 2D Gabor function[14,16,17] (Fig. 3d). Similar results were found for all other columns in our sample ($n = 73$). Fits to the 2D Gabor function show the similarity across columns in the relative phase of the aggregate subfields, number of half-cycles, and aspect ratio (Fig. 3e), and these fits accurately predict each column's preferred orientation and visuotopic location (Fig. 3f). The spatial frequency preference measured with grating stimuli was systematically underestimated by the linear receptive field analysis, consistent with previous observations from electrophysiological recordings[15].

The fact that OFF inputs serve as the anchor for the aggregate receptive fields of cortical columns and that OFF inputs exhibit a precise visuotopic organization predicts that preference for absolute spatial phase—the phase of a sine wave grating stimulus measured relative to a common reference point in visual space—should also be mapped in a smooth and continuous fashion in the responses of layer 2/3 neurons. To test this prediction, we examined the cortical responses of large populations of single neurons while presenting an elongated two-period grating with different phases relative to the centre location of the population receptive field (see Methods). This stimulus revealed a strong preference for absolute spatial phase in the majority of layer 2/3 neurons (83.7% tuned; mean tuning bandwidth ± s.d., 16.9° ± 5.55°) and an orderly progression across the cortical surface (Fig. 4a, b). The phase tuning curve for the neurons can be well described by circular Gaussian curve fitting and is consistent with the receptive field structure of the neurons (Extended Data Fig. 7a, b). The absolute spatial phase parameter of the Gabor fit to the ARF accurately predicts the

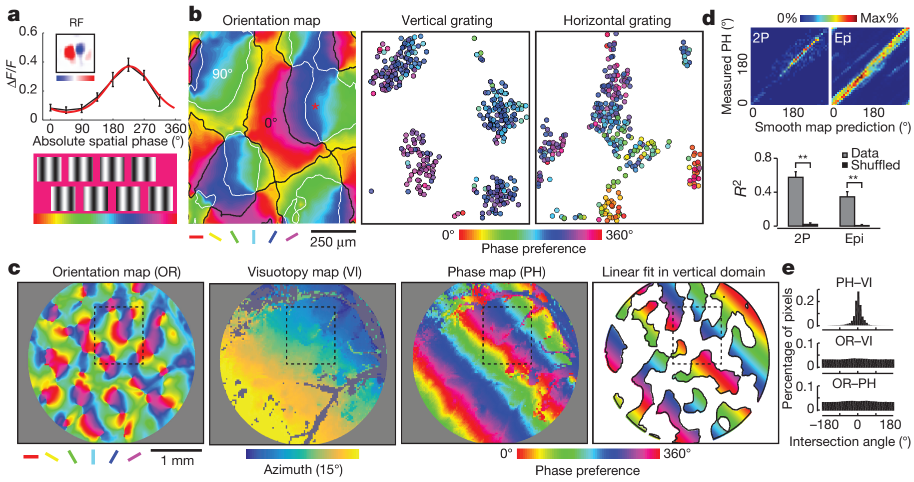


**Figure 4 | Smooth progression of absolute spatial phase across orientation domains.** **a**, The phase tuning curve (black) and its Gaussian fit (red) for an example neuron derived from eight static grating stimuli. **b**, Organization of the phase preference for populations of neurons derived with vertical and horizontal grating stimuli visualized with two-photon imaging at three cortical depths. Cortical domains that are visually responsive (see Methods) to vertical and horizontal gratings are delineated by contours (white and black, respectively). Neighbouring neurons exhibit similar phase preferences, and the preferences shift progressively across the orientation domains. **c**, Epi-fluorescence imaging demonstrates relation of phase map derived with vertical grating to maps of orientation (OR) and visual space (azimuth; VI). Black rectangles indicate the two-photon imaging field of view shown in **b**. The smooth progression of preferred phase (PH) along the visuotopic axis orthogonal to the stimulus orientation is evident at this scale. The rightmost figure shows a linear fit of the phase signal within vertical orientation domains to approximate the phase preference map. **d**, For both the two-photon (2P) and epi-fluorescence (Epi) data, a smooth phase progression generated with a linear fit was used to test for correlation with the experimental data (circular regression, both $P < 0.0001$). The smooth progression accounted for a greater amount of the variance in the experimental data than did the shuffled data ($**P < 0.0001$, rank-sum test within group; see Methods). Error bars indicate s.e.m. **e**, The intersection of the phase and visuotopic map gradients shown in **c** peaks around 0° degree (0.32° in this case and −0.08° on average of six maps), indicating a parallel relationship ($P = 5.2 \times 10^{-16}$, Rayleigh test), while there is no significant non-uniformity for the intersection of orientation map gradients with either phase or visuotopic map gradients ($P > 0.05$, Rayleigh test).

preferred phase of the column (Extended Data Fig. 7c). For a given region of cortex, we found that the preferred phase was comparable at multiple depths, consistent with a columnar organization (Extended Data Fig. 7d, e).

The systematic mapping of preferred absolute spatial phase is especially evident at the larger spatial scales that can be visualized using wide-field epi-fluorescence imaging (Fig. 4c). In these images, the spread of the fluorescent signals beyond the cell bodies of stimulated orientation columns (due to calcium signals in the neuropil as well as light scattering) emphasizes the linear progression of the phase map along the axis of visual space orthogonal to the stimulus orientation. The linear fit to the experimental data for each 360° periodic phase cycle accounts for the organization of the absolute spatial phase preference map (Fig. 4d) and the intersection angle between the gradient of the maps for visuotopy and absolute spatial phase shows that they are parallel to each other (Fig. 4e and Extended Data Fig. 7f, g). Moreover, our analysis indicates that the cortical area corresponding to 1° of visual space contains complete coverage for phase and orientation, and that the uniformity of coverage increases with the amount of visual space included (Extended Data Fig. 8).

The striking differences between the topologies of ON and OFF inputs to layer 2/3 simple cells is reminiscent of the structural and functional differences that have been described for ON- and OFF-centre retinal ganglion cells in a number of species[18–20]. OFF-centre retinal ganglion cells are more numerous and have smaller dendritic fields than ON-centre retinal ganglion cells, endowing them with a capacity for greater spatial resolution that is consistent with the presence of more regions of negative than positive contrast in natural scenes[20]. Previous computational[21,22] and experimental studies[3,4] have recognized that the spatial arrangement of ON- and OFF-centre inputs is likely to provide the scaffold for orientation column structure, but exactly how the ON and OFF pathways converge to generate coherent maps of orientation and visual space has remained unclear. Our results provide evidence that the retinal asymmetries in the ON and OFF pathways are reflected in cortical map structure such that simple cell receptive fields preserve a high degree of visuotopic order in their OFF subfield inputs, while exploiting visuotopic displacement of their ON subfield inputs to generate an orderly representation of orientation preference. In addition, the resulting OFF-anchored columnar architecture enables emergence of an orderly representation of absolute spatial phase—a property that contains a wealth of information about the visual scene that can be used to efficiently encode spatial patterns[23,24], motion[25], and depth[26]. We emphasize that the modular representation of absolute spatial phase preference demonstrated here is distinct from the modular representation of polarity preference that has been described in cat[3,4] and ferret[27] visual cortex. Although there are species differences in the representation of polarity, electrophysiological studies in layer 4 of the cat visual cortex suggest that common rules govern the convergence of ON and OFF inputs to build orientation-selective simple cells, and that the modular representation of absolute spatial phase is a general principle of cortical organization that is common to a broad range of species with well-developed columnar architecture[28]. Interestingly, despite the local diversity of orientation preference and receptive field structure in mouse visual cortex, adjacent neurons exhibit specificity in the overlap of their ON and/or OFF subfields that is predictive of connectivity[29,30], consistent with the idea that the topology of ON and OFF inputs shapes the organization of cortical circuits even in the absence of cortical columns.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 11 October 2015; accepted 21 March 2016.**
**Published online 27 April 2016.**

1. Hubel, D. H. & Wiesel, T. N. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol. (Lond.)* **160,** 106–154 (1962).
2. Ferster, D., Chung, S. & Wheat, H. Orientation selectivity of thalamic input to simple cells of cat visual cortex. *Nature* **380,** 249–252 (1996).

3. Jin, J. Z. *et al.* On and off domains of geniculate afferents in cat primary visual cortex. *Nature Neurosci.* **11,** 88–94 (2008).
4. Jin, J., Wang, Y., Swadlow, H. A. & Alonso, J. M. Population receptive fields of ON and OFF thalamic inputs to an orientation column in visual cortex. *Nature Neurosci.* **14,** 232–238 (2011).
5. Bosking, W. H., Crowley, J. C. & Fitzpatrick, D. Spatial coding of position and orientation in primary visual cortex. *Nature Neurosci.* **5,** 874–882 (2002).
6. Reid, R. C. & Alonso, J. M. Specificity of monosynaptic connections from thalamus to visual cortex. *Nature* **378,** 281–284 (1995).
7. Alonso, J. M., Usrey, W. M. & Reid, R. C. Rules of connectivity between geniculate cells and simple cells in cat primary visual cortex. *J. Neurosci.* **21,** 4002–4015 (2001).
8. Fan, Y. *et al.* Genome of the Chinese tree shrew. *Nat. Commun.* **4,** 1426 (2013).
9. Rockland, K. S. & Lund, J. S. Widespread periodic intrinsic connections in tree shrew visual cortex (Area 17). *Science* **215,** 1532–1534 (1982).
10. Bosking, W. H., Zhang, Y., Schofield, B. & Fitzpatrick, D. Orientation selectivity and the arrangement of horizontal connections in tree shrew striate cortex. *J. Neurosci.* **17,** 2112–2127 (1997).
11. Van Hooser, S. D. *et al.* Transformation of receptive field properties from lateral geniculate nucleus to superficial V1 in the tree shrew. *J. Neurosci.* **33,** 11494–11505 (2013).
12. Kretz, R., Rager, G. & Norton, T. T. Laminar organization of ON and OFF regions and ocular dominance in the striate cortex of the tree shrew (*Tupaia belangeri*). *J. Comp. Neurol.* **251,** 135–145 (1986).
13. Muly, E. C. & Fitzpatrick, D. The morphological basis for binocular and ON/OFF convergence in tree shrew striate cortex. *J. Neurosci.* **12,** 1319–1334 (1992).
14. Jones, J. P. & Palmer, L. A. The two-dimensional spatial structure of simple receptive fields in cat striate cortex. *J. Neurophysiol.* **58,** 1187–1211 (1987).
15. Lampl, I., Anderson, J. S., Gillespie, D. C. & Ferster, D. Prediction of orientation selectivity from receptive field architecture in simple cells of cat visual cortex. *Neuron* **30,** 263–274 (2001).
16. Ringach, D. L. Spatial structure and symmetry of simple-cell receptive fields in macaque primary visual cortex. *J. Neurophysiol.* **88,** 455–463 (2002).
17. Yeh, C. I., Xing, D., Williams, P. E. & Shapley, R. M. Stimulus ensemble and cortical layer determine V1 spatial receptive fields. *Proc. Natl Acad. Sci. USA* **106,** 14652–14657 (2009).
18. Chichilnisky, E. J. & Kalmar, R. S. Functional asymmetries in ON and OFF ganglion cells of primate retina. *J. Neurosci.* **22,** 2737–2747 (2002).
19. Devries, S. H. & Baylor, D. A. Mosaic arrangement of ganglion cell receptive fields in rabbit retina. *J. Neurophysiol.* **78,** 2048–2060 (1997).
20. Ratliff, C. P., Borghuis, B. G., Kao, Y. H., Sterling, P. & Balasubramanian, V. Retina is structured to process an excess of darkness in natural scenes. *Proc. Natl Acad. Sci. USA* **107,** 17368–17373 (2010).
21. Miller, K. D. A model for the development of simple cell receptive fields and the ordered arrangement of orientation columns through activity-dependent competition between ON- and OFF-center inputs. *J. Neurosci.* **14,** 409–441 (1994).
22. Ringach, D. L. Haphazard wiring of simple receptive fields and orientation columns in visual cortex. *J. Neurophysiol.* **92,** 468–476 (2004).
23. Oppenheim, A. V. & Lim, J. S. The importance of phase in signals. *Proc. IEEE* **69,** 529–541 (1981).
24. Pollen, D. A. & Ronner, S. F. Phase relationships between adjacent simple cells in the visual cortex. *Science* **212,** 1409–1411 (1981).
25. Adelson, E. H. & Bergen, J. R. Spatiotemporal energy models for the perception of motion. *J. Opt. Soc. Am. A* **2,** 284–299 (1985).
26. Kara, P. & Boyd, J. D. A micro-architecture for binocular disparity and ocular dominance in visual cortex. *Nature* **458,** 627–631 (2009).
27. Smith, G. B., Whitney, D. E. & Fitzpatrick, D. Modular representation of luminance polarity in the superficial layers of primary visual cortex. *Neuron* **88,** 805–818 (2015).
28. Wang, Y. *et al.* Columnar organization of spatial phase in visual cortex. *Nature Neurosci.* **18,** 97–103 (2015).
29. Smith, S. L. & Häusser, M. Parallel processing of visual space by neighboring neurons in mouse visual cortex. *Nature Neurosci.* **13,** 1144–1149 (2010).
30. Cossell, L. *et al.* Functional organization of excitatory synaptic strength in primary visual cortex. *Nature* **518,** 399–403 (2015).


**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** We thank members of the Fitzpatrick laboratory for discussions and comments, D. Ouimet and K. Diah for animal technical support, T. Walker for technical assistance, A. Jacob for histology assistance, D. Wilson for advice on two-photon imaging, and D. Whitney for advice on epi-fluorescence imaging. This work was supported by grants from the US National Institutes of Health and funding from Max Planck Florida Institute for Neuroscience and Max Planck Society to D.F.

**Author Contributions** K.-S.L. and D.F. designed the experiments. X.H. helped to collect data in the initial stage. K.-S.L performed all the experiments and analysed all the data with assistance from X.H. K.-S.L. and D.F. wrote the manuscript with input from X.H.

**Author Information** Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to D.F. (david.fitzpatrick@mpfi.org).

## METHODS

All experimental procedures were approved by the Max Planck Florida Institute for Neuroscience Animal Care and Use Committee and performed in compliance with guidelines published by the National Institutes of Health. Tree shrews (*Tupaia belangeri*, $n = 18$, 2–4 months of age, male and female) were injected with a virus expressing GCaMP6s[31], and then used in a terminal imaging experiment after a 10–15-day survival period. Animal numbers were minimized to conform to ethical guidelines while accurately measuring parameters of animal physiology. No statistical methods were used to predetermine sample size.

**Viral expression of GCaMP6s.** Tree shrews were initially anaesthetized with Midazolam ($100\,mg\,kg^{-1}$, intramuscular (IM)) and ketamine ($100\,mg\,kg^{-1}$, IM) and given atropine ($0.5\,mg\,kg^{-1}$, subcutaneous (SC)) to reduce secretions. A long-acting analgesic (slow-release Buprenorphine, $0.6\,mg\,kg^{-1}$, SC) was administered before the surgery. The animal's head was shaved, any remaining hair was removed with Nair, and the surgical site was injected with a mixture of bupivacaine and lidocaine (0.3–0.5 ml, SC). A mixture of oxygen and nitrous oxide ($O_2/N_2O$ 1:0 to 1:2) and gas anaesthesia (isoflurane 0.5–2%) were initially delivered through a mask and later through an intubation tube. Venous cannulation (tail or hind limb) and tracheal intubation were established after the animal no longer responded to toe-pinching. Internal temperature (37–38 °C) was maintained by a thermostatically controlled heating pad while expired $CO_2$ and heart rate were monitored for signs of stress. Artificial respiration was provided at between 100 and 130 strokes per minute through a ventilator. The animal was placed in a stereotaxic device (Kopf, Model 900 Small Animal Stereotaxic Instrument), a small incision was made, skin and muscle were retracted, and a small craniotomy (about 1-mm diameter) was made over the centre of the primary visual cortex. The visual cortex was injected with a total of 1–2 μl of virus solution ($1 \times 10^{13}$ GC ml$^{-1}$ to $2 \times 10^{13}$ GC ml$^{-1}$) containing AAV2/9-Syn-GCaMP6s.WPRE.SV40 (Penn Vector Core) through a bevelled glass micropipette (tip size 10–20 μm diameter, Drummond Scientific Company) using a nanoinjector (Drummond Nanoject II, WPI). Only one injection site was placed in the cortex of each animal. To facilitate spreading, the virus was delivered at two depths, 200 and 400 μm below the cortical surface. After the injection, the craniotomy was covered with bone wax and the scalp incision was closed with 4-0 Ethilon sutures. Neosporin was applied to the wound margins. The animals were then placed on a heating blanket in a small cage to recover from anaesthesia. A period of 10–15 days was allowed for expression time before two-photon imaging experiments were carried out. GCaMP6s expression was found in approximately $84 \pm 2.3\%$ of the neurons in the superficial part of layer 2/3 (within 500 μm of the surface, $n = 4$ animals). The densely labelled area was generally 5–7 mm in diameter.

**Preparation for two-photon imaging experiments.** After 10–15 days of expression, anaesthesia was induced with Midazolam ($100\,mg\,kg^{-1}$, IM) and ketamine ($100\,mg\,kg^{-1}$ and $0.2$–$0.5\,mg\,kg^{-1}$, IM), and atropine ($0.5\,mg\,kg^{-1}$, SC) was given to reduce secretions. An analgesic (buprenorphine, $0.3$–$0.6\,mg\,kg^{-1}$, SC) was administered before the surgery. A peripheral venous line on either tail or hind limbs was prepared for delivering fluid during surgery and muscle relaxants during imaging experiments. All regions identified as incision sites for the surgery were treated with a mixture of bupivacaine and lidocaine (0.3–0.5 ml, SC), and ear bars were coated with Lidocaine ointment (5%). Gas anaesthesia (isoflurane 0.5–2% in $O_2/N_2O$ 1:0 to 1:2) was delivered via artificial respiration following intubation or tracheotomy. The animal's head was shaved and placed in a customized stereotaxic device that did not obstruct the view of the stimulus screen. Body temperature (37–38 °C) was maintained by a thermostatically controlled heating pad and expired $CO_2$ (3.5–4.5%) and heart rate were monitored for signs of stress. A roughly 2-cm incision was made over the skull near the midline, and the skin and muscle were retracted. A head-plate with a central opening was attached to the skull with dental cement (C & B Metabond) and a craniotomy ($6 \times 6$ mm) was made, centred over the injection site of GCaMP6s within primary visual cortex. After the dura mater was removed, a piece of double-layer cover slip composed of a small round glass coverslip (3 mm diameter, 0.7 mm thickness, Warner Instruments) glued to a larger coverslip (8 mm diameter, 0.17 mm thickness, Electron Microscopy Sciences) with an optical adhesive (Norland Optical Adhensive 71) was placed onto the brain to gently compress the cortex and reduce biological motion during imaging. For some animals, a custom metal insert (5 mm diameter, 0.5 mm thickness for inner hole) attached with a coverslip (5 mm diameter, 0.17 mm thickness, Electron Microscopy Sciences) was used to achieve better optical transmission and a larger field of view. The cover slip or metal insert was sealed with a snap ring (5/16-inch internal retaining ring, McMaster-Carr) that fit into the chamber. Contact lenses were placed on both eyes for protection and stability. During imaging experiments, the isoflurane level was decreased to 0.5–1%. Pancuronium bromide or vecuronium bromide ($2\,mg\,kg^{-1}\,h^{-1}$, intravenous (IV)) was used as a paralytic to prevent eye movements.

**Two-photon imaging experiments.** Imaging experiments were performed using a B-Scope (Thorlabs) with either 910 nm excitation provided by an InSight DS+ (Spectra-Physics) or 910 nm excitation provided by a Mai Tai DeepSee laser (Spectra-Physics), running Scanimage 4.1 or 4.2 (Vidrio Technologies)[32]. Average excitation power at the exit of the objective (16×, CFI75, Nikon Instruments) ranged from 16 to 40 mW. Images were acquired at 15–30 Hz ($512 \times 512$ pixels, field of view (FOV) ranges from $0.44 \times 0.44\,mm^2$ to $1.1 \times 1.1\,mm^2$). Two-photon frame triggers from Scanimage and events denoting stimulus onset, stimulus offset, and stimulus identity were recorded using Spike2 (CED; Cambridge, UK). In a typical imaging session lasting about 16 h, 2–4 different fields of view were sampled and, at each site, data were acquired at 2–4 different depths with at least 35 μm separation, between 50 and 350 μm below the cortical surface. *Z*-stacks of individual fields of view were acquired by averaging 50 frames per plane using 1-μm steps from the surface to about 350 μm deep.

**Epi-fluorescence imaging experiments.** Epi-fluorescence imaging was performed using a custom light path on the B-Scope with 525 nm LED illumination (Thorlabs). GCaMP6s fluorescence signal from the cortical surface was acquired at about 15 Hz ($640 \times 540$ pixels, field of view (FOV) ranges from $3 \times 2.53\,mm^2$ to $4 \times 3.38\,mm^2$) using a Xyla sCMOS camera (Andor) controlled by μManager2. Average excitation power at the exit of the objective (4×, UPlanFl, Olympus) ranged from 0.2 to 0.8 mW. Epi-fluorescence frame triggers from μManager2 and stimulus events were recorded using Spike2 (CED; Cambridge, UK). The visual stimulus and the analysis method were the same as two-photon pixel-based experimental design. *Z*-projections of two-photon fields of view were aligned to the epi-fluorescence imaging with the blood vessel pattern.

**Visual stimulation.** Visual stimuli were displayed on an LED monitor (29 cm (height) × 51 cm (width)) with a resolution of 1,920 × 1,080 pixels, which was placed in front of the centre of the animal to cover about 100° in azimuth and 70° in elevation. The refresh rate of the monitor was 120 Hz, and the mean luminance for grey background was $54\,cd\,m^{-2}$. The stimulus monitor was placed at a distance of 21 cm from the eyes. Receptive fields of neurons in the field of view usually appeared close to the centre of the monitor. Visual stimuli were generated using Psychopy2 written in Python. There were two main types of visual stimulation experiment: 1) sparse noise stimulus for mapping the receptive field[14]; and 2) grating or bar stimulus for accessing tuning properties. Sparse noise stimuli were composed of two non-overlapped squares ($2 \times 2$ grid size, separated within a 4-grid unit, with black or white sign presented independently) on a $17 \times 17$ square grid grey background (a total of 7,904 images), which occupied 17 or 25° of visual space. Individual images were presented for 200 ms without an inter-stimulus interval. Depending on imaging quality, 1–2 trials of the entire stimulus set was presented, which lasted 26–52 min. To measure orientation tuning properties, square wave gratings (contrast 100%, spatial frequency (SF) 0.25 cycles per degree (CPD), and temporal frequency (TF) 4 Hz, stimulus duration 2 s, full screen) drifting in both directions were presented at 16 different orientations (0–168.75°, spaced at 11.25°). To measure visuotopic position tuning, a single static bar (either black or white with 100% contrast, 6° by 24°, 1.5-s duration, centred on the population receptive field centre) was presented at eight different positions (0–21°, spaced at 3°). To measure spatial frequency tuning, sine wave gratings (contrast 100%, preferred orientation for the recording site, and temporal frequency 4 Hz, stimuli duration 2 s, full screen) drifting in both directions were presented at eight different spatial frequencies (0.025–3.2 CPD, spaced at 1 unit in $\log_2$ scale). To measure absolute spatial phase tuning, static sine wave gratings (preferred orientation, contrast 100%, size 15° by 60°, and spatial frequency 0.25–0.35 CPD) were centred on the population receptive field and were presented at eight different phases (0–315°, spaced at 45°) for 1.5 s. Typically 10 stimulus trials (for orientation and spatial frequency) or 20 stimulus trials (for visuotopy and phase) were presented along with blank stimulus trials (random order) with 2–5-s inter-stimulus intervals.

**Perfusion and Histology.** At the end of the experiments, the animal received a lethal injection of Euthasol and was perfused transcardially with 0.9% saline, followed by 4% paraformaldehyde or 10% formalin. The brain was then removed, post-fixed in 4% paraformaldehyde or 10% formalin overnight, transferred to a 30% sucrose solution in phosphate buffer (PB, pH 6.8) and stored at 4 °C for at least two days. The area of interest was blocked and then cut on a freezing microtome with 50-μm-thick parasagittal, coronal or tangential sections collected in serial order. For immunostaining, the slices were incubated in blocking solution for 30 min and then transferred to the primary antibody solution (chicken anti-GFP, 1:1,000, Aves Labs, GFP-1020; rabbit anti-NeuN, 1:1,000, Millipore, ABN78; guinea pig anti-vGlut2, 1:10,000, Millipore, AB2251; mouse anti-PV, 1:1,000, Swant; 235) for overnight incubation at 4 °C. The slices were then incubated in secondary antibody (Alexa 405 for NeuN, 488 for GFP, 568 for vGlut2, 647 for PV; Invitrogen) for 2 h at room temperature, mounted on glass slides, dried, and coverslipped. Labelled neurons and structures were viewed on a fluorescence microscope

(Olympus BX53) or confocal microscope (Zeiss 710 Confocal, 20× objective). We verified that all the injection and imaging sites were centred within 1 mm from the centre of the primary visual cortex, which processes visual information covering approximately −10 to 10° in elevation from the horizontal meridian and 0 to 20° in azimuth from the vertical meridian.

**Data analysis.** Mechanical drift in the imaging plane was corrected using a customized motion registration program written in Matlab (Mathworks). Analyses were performed using custom code written in Matlab or Java package for running ImageJ within Matlab (Miji)[33]. The circular regions of interest (ROIs) corresponding to visually identified neurons were selected using ImageJ. ROIs were drawn by hand and selected by viewing the average intensity or standard deviation z-projection of the stack of two-photon images from one experiment, combined with visual examination of individual imaging frames. The fluorescence of each cell was measured by averaging all pixels within the ROI. Whenever we probed the feature selectivity of a column, we always first aligned and collapsed multiple cortical depths (at least three) from the same two-photon field of view into a 2D field of view for straightforward visualization.

**Sparse noise receptive field measurement.** Hand-mapping with a customized program written in Psychopy2 (ref. 34) was used to determine the area of visual space relevant for the cortical field of view and to verify the stability of the population receptive field during the experimental session. The sparse noise stimulus was then centred on the population receptive field centre, ensuring that all of the neurons in the imaged area had receptive fields that fell within the stimulus presentation area. Linear receptive fields (RF) were obtained by reverse-correlating neuronal responses to an image set containing both white and black squares. The distributions of ON and OFF response regions were obtained by reverse correlation to image sets containing a single contrast polarity (either white or black, respectively).

Reverse correlation analysis began by filtering the fluorescence signal from individual neurons (7 Hz low-pass zero-phase) and then applying a threshold at 2 s.d. from the mean. Individual peaks in the trace were detected as fluorescence events, and the area under the curve of the rising phase was assigned to the peak time as a measure of the response strength for the corresponding fluorescence event[35]. The spatiotemporal receptive field was reconstructed with repeating reverse correlation from 0 to 850 ms after stimulus onset in 50-ms time windows. The signal-to-noise ratio (SNR) for time $\tau$ was calculated as the ratio of the spatial variance between time $\tau$ and the stimulus onset time:

$$\sigma_{xy}^2(\tau)/\sigma_{xy}^2(\text{stimulus onset}),\ \text{where } \sigma_{xy}^2(\tau) = [R(x,y,\tau) - \text{mean}(R(x,y,\tau))]_{x,y}^2$$

The peak SNR for each neuron was obtained from a smooth spline fitting of the SNR curve. Receptive fields or individual ON or OFF response regions of neurons whose peak SNR values did not exceed 2.2 were not included in the analysis, thereby excluding spontaneously active, non-responsive cells. The receptive field was then resized to 48 × 48 pixels and the significance of each pixel was assessed by comparison with pixels from 100 randomly shuffled receptive fields (reverse correlation repeated 100 times for each neuron, using the time of the peak SNR). A significant pixel was defined as a pixel with an absolute value higher than the absolute value of the mean + 5 s.d. of the shuffled receptive field. Neurons whose receptive fields did not contain any significant pixels were excluded from further analysis. Fifty seven percent of the neurons that were imaged met both the signal-to-noise and pixel statistical significance criteria, and were used for further analyses. To verify that our method for analysing calcium signals did not impact the structure of the RFs, we selected RFs with high SNR (>10) and computed the similarity index (SI)[17] of RFs derived from our original method with RFs derived from six alternative methods:

$$\frac{\sum_{x,y} \text{RF}_0(x,y)\text{RF}_1(x,y)}{\sqrt{\sum_{x,y} \text{RF}_0^2(x,y)\text{RF}_1^2(x,y)}}$$

where $\text{RF}_0$ is the RF derived with the original method describing above, and $\text{RF}_1$ the RF derived with an alternative method. The alternative methods included RFs derived: (1) without applying a filter to the calcium trace; (2) without applying a threshold of two standard deviations; (3) with the response strength of each event calculated as the area under the curve (AUC) of each peak of the fluorescence signal or (4) as AUC of each peak of fluorescence signal relative to the baseline fluorescence level; (5) with event time assigned to the initial deflection point of each fluorescence peak; and (6) with event detection performed using a standard de-convolution method[36].

Neurons were placed into one of three classes: simple, complex, and single sign (either ON or OFF). Forty-nine per cent of the neurons were found to have statistically significant responses to only one sign of the sparse noise stimulus (either ON or OFF) and were categorized as single sign. Others with statistically significant responses to both dark and light sparse noise stimuli were further characterized as simple or complex cells based on the degree of segregation of the ON and OFF response fields. For this purpose we used an ON/OFF segregation (seg) index:

$$\text{ON/OFF seg} = \frac{\sum_p |R'\text{ON}(p) - R'\text{OFF}(p)|}{\sum_p R'\text{ON}(p) + R'\text{OFF}(p)}$$

where $R'\text{ON}(p)$ is all the pixels modulated by the light sparse noise stimulus and $R'\text{OFF}(p)$ is all the pixels modulated by the dark sparse noise stimulus. Neurons with an ON/OFF segregation index greater than 0.6 were classified as simple cells while those with an index less than 0.6 were classified as complex cells[11]. For those neurons that responded significantly to both dark and light sparse noise stimuli, we also characterized the relative effectiveness of dark and light stimuli by calculating the ON/OFF ratio:

$$\text{ON/OFF ratio} = \frac{\max(\text{SNR}_{\text{ON}})}{\max(\text{SNR}_{\text{ON}}) + \max(\text{SNR}_{\text{OFF}})}$$

To evaluate the visuotopic organization of receptive field centres, and the centres of the ON and OFF subfields of simple cells, we calculated the centres of mass of these regions in visual space using the absolute value of all the significant pixels in the receptive field or the subfield, respectively. In a few cases in which a simple cell had more than one significant ON or OFF subfield (16%), only the subfield with the maximum (for ON) or minimum (for OFF) value was selected for estimating the neuron's subfield centre of mass. Thus each simple cell could be summarized as having one receptive field centre, one ON subfield centre and one OFF subfield centre.

To test whether the offset between the ON and OFF subfields of simple cells could be used to estimate the cell's preferred orientation, the predicted orientation preference was defined as the orientation perpendicular to the axis of the receptive field dipole (a line connecting the ON and OFF centres). To test whether the phase tuning can be predicted from the simple cell receptive field structure, the receptive field was used as a 2D filter and convolved with a stimulus grating with eight phases to derive a predicted phase tuning curve, which was then fit with a Gaussian. This result was then compared with the experimental phase measurement data.

To characterize the feature selectivity of simple cells in a single orientation column, we first aligned and collapsed all the cortical depths (at least three) from the same two-photon field of view into a 2D neuron population. To sample from columns of cells with similar orientation preference, we employed an 80-μm diameter sample window and moved it in 5-μm steps across the field of view, searching for sites that met the following criteria: more than 12 simple cells with a maximum orientation difference from the mean of the population less than 11.25° and not containing more than four cells that were included in other columns. The values that we used to define an orientation column were based on the average bandwidth of the active zones produced in cortex by the presentation of a single grating stimulus. In cortical distance, the mean full width at half maximum was $86.7 \pm 7.6\,\mu\text{m}$ ($n = 106$). On average, the range of orientation preferences exhibited by the neurons in this cortical area extended $11.9 \pm 2.4°$ beyond the orientation of the stimulus half width at half maximum (HWHM). The ARF was then computed as the average of the normalized amplitude of all the simple cells within the orientation column. To estimate the predicted feature selectivity from the ARF, the ARF was fitted and parameterized with a two-dimensional Gabor function using the Levenberg–Marquardt algorithm[37]. The Gabor function is described by

$$G(x,y) = A\exp\left(-\frac{x'^2}{2\sigma_x^2} - \frac{y'^2}{2\sigma_y^2}\right)\cos(2\pi f x' + \varphi)$$

where $(x', y')$ is obtained by translating the original coordinate system and rotating it by $\theta$:

$$x' = (x - c_x)\cos\theta - (y - c_y)\sin\theta$$

$$y' = (x - c_x)\sin\theta + (y - c_y)\cos\theta$$

The Gabor function can be also viewed as an underlying 2D cosine grating parameterized by $\theta$ (orientation), $f$ (spatial frequency) and $\varphi$ (relative spatial phase), which is enveloped by a 2D Gaussian function parameterized by $A$ (amplitude), $c_x$ and $c_y$ (centre of the Gaussian), and $\sigma_x$ and $\sigma_y$ (standard deviations of the

Gaussian in perpendicular and parallel axis of the grating, respectively). All the fractions of explained variance from Gabor fits were at least 0.7. The aspect ratio of the Gabor fit was defined as $\sigma_y/\sigma_x$ and the number of half-cycles within the Gaussian envelop was defined as $8f\sigma_x$. To assess the preference of the orientation column from the experimental tuning data, we used the average (or circular average for orientation and absolute spatial phase) of the preferences of all of the simple cells within the column. To evaluate the relation between orientation preference and the angle of ON subfield displacement, the ON-dipoles for the orientation column were determined by applying *k*-means clustering (Matlab) to all the ON centres within the column with the assumption of two clusters. The predicted orientation preference for the column was defined as the orientation orthogonal to the axis of the ON-dipoles.

**Tuning curve and preference map for multiple visual properties.** For computing tuning properties, the fluorescence signal was calculated as $\Delta F/F = (F - F_0)/F_0$, where $F_0$ is the baseline fluorescence signal averaged over a 1-s period immediately before the start of visual stimulation, and $F$ is the fluorescence signal averaged over the first 1.5-s period after the start of the visual stimulation. For example, orientation tuning curves were obtained by calculating the mean fluorescence signal ($\Delta F/F$) for each orientation, and then fitting a Gaussian curve to the resulting data. Neurons were considered to be visually responsive if the maximum stimulus-related fluorescence response ($\Delta F/F$) to any orientation was greater than 5% on average, and also greater than 2 s.d. above the mean baseline fluorescence. In addition, we required that cells respond at least 2 s.d. above baseline on at least 20% of the trials tested. Neurons were considered to be orientation tuned if they were visually responsive and also met the following criteria: (1) well fit by the Gaussian function ($r > 0.7$, $P < 0.05$), and (2) tuning index (TI) $> 0.4$

$$\mathrm{TI} = \frac{\mu_{\mathrm{pref}} - \mu_{\mathrm{ortho}}}{\mu_{\mathrm{pref}} + \mu_{\mathrm{ortho}}}$$

where $\mu_{\mathrm{pref}}$ equals the mean response to the preferred orientation and $\mu_{\mathrm{ortho}}$ equals the mean response to the orthogonal orientation. For analysis of visuotopy and spatial frequency, a similar index was used but the response to the orthogonal stimulus was replaced by the response to a bar stimulus presented outside population receptive field or response to a grating with spatial frequency at 3.2 CPD, respectively. The preferred tuning properties and tuning width were calculated from the Gaussian curve fitting. For pixel-based preference maps in both two-photon and epi-fluorescence imaging, we used data binned from areas of $10 \times 10$ pixels and assigned the preferred tuning value to each unit followed by smoothing with a 20-μm radius Gaussian filter.

To examine the relationship between the precision of the visuotopic arrangement of ON and OFF subfields and map structure, we computed the local heterogeneity for each cell in the orientation preference map. To obtain local heterogeneity we calculated the circular variance of the orientation tuning distribution of all of the pixels surrounding the cell, weighting the values obtained from each pixel using a Gaussian function with a $\sigma$ of 30 μm and a cut-off at 50 μm:

$$\text{Circular variance} = 1 - \left| \frac{\sum_k r_k \exp(i2q_k)}{\sum_k r_k} \right|$$

Where $r_k$ is the magnitude of the responses ($\Delta F/F$) to the stimulus $k$ with orientation $\theta_k$.

To evaluate the extent of cortical activity evoked by a single bright (ON) or dark (OFF) bar with epi-fluorescence imaging, the $\Delta F/F$ of each pixel in the area of activation was calculated. The width of the activated region was estimated by averaging the response along the visuotopic axis orthogonal to the orientation of the bar stimulus. The HWHM was computed for each visual stimulus, by using stimulus positions in which both the ON and OFF evoked responses were centred within the imaging window.

To evaluate the distribution in visual space of the ON and OFF subfields of simple cell populations, we calculated the pairwise distance between the centres of individual subfields for all ON centres, all OFF centres, all centres with the same signs, and all centres with different signs. These values were compared to the pairwise distances derived from the random shuffling of sign identity in 10 independent repetitions. A similar analysis was performed for RF centres by comparing the pairwise distance of the original RF centres to the pairwise distances created by shuffling the angle in polar coordinates, but keeping the radian from the population centre the same, in 10 independent repetitions. For each imaging area, the pattern of the distribution for any relationship can be summarized as the log ratio between the mean pairwise distance from data and shuffle: a negative value indicates more clustered than by random chance, and a positive value means more scattered than by random chance.

The cellular precision of fine visuotopy was tested for the RF centres and the centres of OFF and ON subfields from the same population of neurons. Elevation and azimuth coordinates were used to evaluate visuotopic precision in each dimension. To quantify deviations from smooth visuotopy, Pearson's correlation coefficients were applied to each visuotopic centre position and their distance along the cortical representation for each visuotopic axis. For each field of view, the average deviation from a smooth visuotopy and the goodness of fit were used as a measure of precision in visuotopy.

The precision of mapping for absolute spatial phase was tested at the cellular level with two-photon imaging and over a larger scale with epi-fluorescence imaging. Because absolute spatial phase maps are defined with a static grating of the same orientation and the full range of phases, we calculated the smoothness of the phase map by limiting analysis to regions of cortex responsive to the testing orientation. To define the single orientation-responsive region (SORG) for constraining the phase map, orientation contour lines were drawn for the specific testing orientation $\pm 30°$ on the filtered orientation preference map. This is based on the average size of the responsive area evoked by flashing a static grating stimulus with a single orientation under two-photon imaging. Neurons or pixels falling within the orientation preference contour lines of SORG were then selected for further analysis. We generated a theoretical prediction of a smooth phase map by projecting the 2D phase map along the orthogonal axis of visuotopy and applied a linear fit to each 360° periodic cycle. The 1D phase gradient was then transformed into a 2D phase map with interpolation. The phase preferences of individual cells from two-photon imaging or those of individual pixels in epi-fluorescence imaging were then correlated to the prediction from the smooth phase map. To measure the intersection angle between maps[38], the vector indicating the change in the preferred feature was extracted for each pixel, and the angular difference of all the pixels within the region of interest was compared for different functional maps. To characterize the statistical structure of functional maps, we compared the relationship between Δcortical distance and the mean of Δpreferred feature. The cortical distance shorter than the first agreement between the real and shuffled data was defined as the clustering effect of the preferred feature. The periodicity of the map was computed by sinusoidal fit to the data points beyond the distance of feature clustering.

**General statistics.** Statistical analyses were performed in Matlab. We used a two-sided non-parametric Wilcoxon rank-sum test to compare two groups and the Kruskal–Wallis test to compare multiple groups with post-hoc tests using Dunn's test, without assumptions of normality or equal variances. Circular correlation coefficient was used for orientation and spatial phase, while Pearson's correlation coefficient was applied to visuotopy and spatial frequency. The Rayleigh test was used to test the uniformity of the intersection angle distribution between the two maps. All statistical methods were two-sided. No estimates of statistical power were performed before experiments.

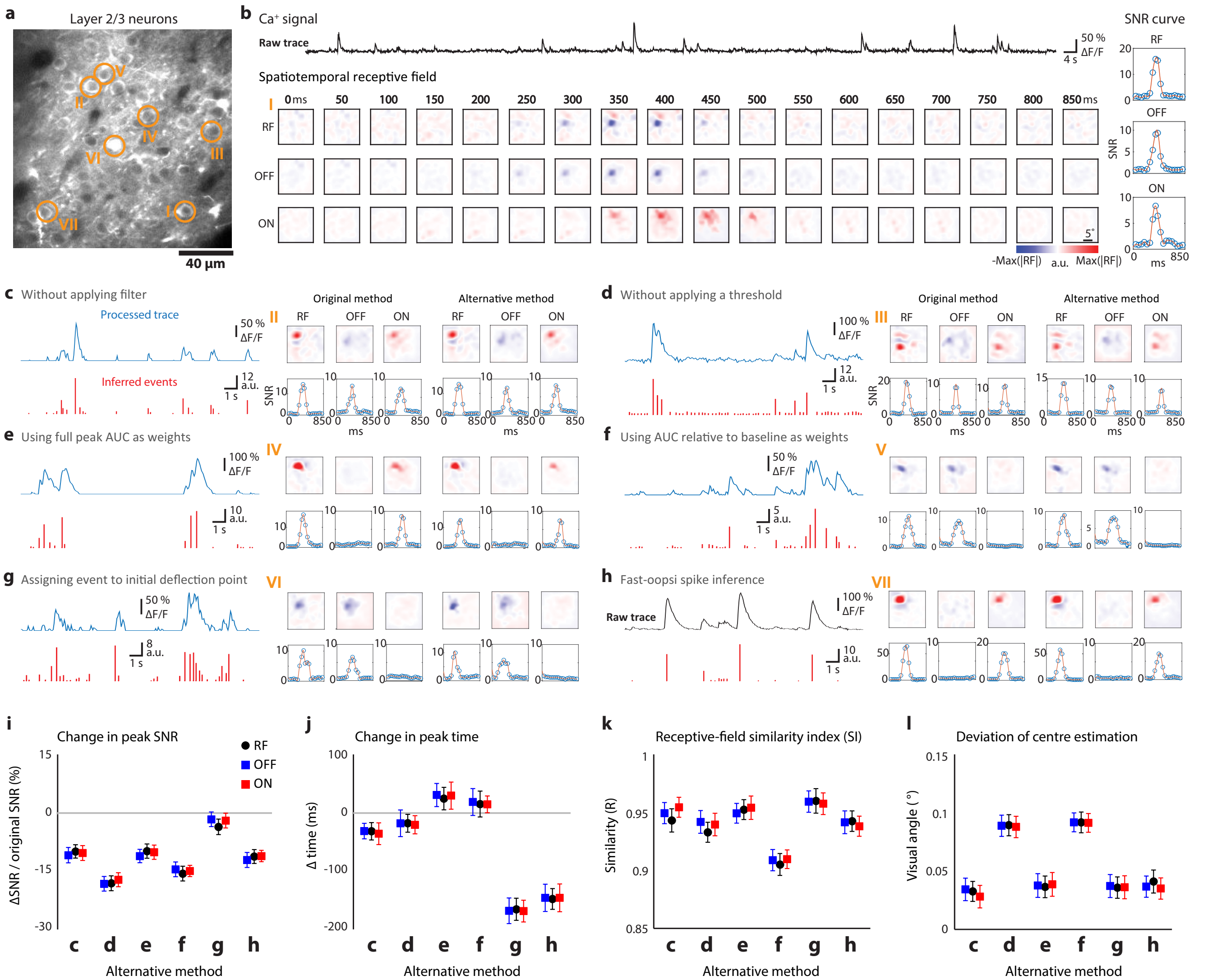
**V1 computational model.** To evaluate the uniformity and completeness of coverage for orientation and absolute spatial phase, we began by simulating the underlying receptive field structures from the large-scale functional maps. We used the orientation map derived from intrinsic signal imaging in the tree shrew[39] and the phase map, generated using published data on visuotopy and cortical magnification factor[5,10]. We then generated a Gabor-like simple cell receptive field for each pixel, using the typical scale for single cortical neurons. The organization of ON and OFF subfields for each pixel was then modelled using the relationships described in the results of this study: that is, the centres of OFF subfields of the cortical population followed a perfect visuotopic map, distributed according to the cortical magnification factor, whereas the centres of ON subfields followed an orientation-specific displacement for the simple cell, randomly placed on either side of the OFF centre with a distance of 3.5° visual angle.

We used the receptive field derived from each pixel in the functional map as a spatial filter for visual stimulation. A theoretical cortical response was derived using a circular patch stimulus with both negative and positive contrast, varying in both size and visual location. The output of the cortical activity pattern was then transformed into a luminance scale overlaid on the functional map for visualizing the area and functional properties that were covered. We defined $\delta$ as the distribution of the functional properties covered within the responsive area divided into eight bins for either orientation or spatial phase, and then evaluated the coverage from these distributions in several ways[5,40]: first, the completeness of coverage was calculated as the number of bins with a positive number divided by total numbers of bins. Second, the uniformity of coverage ($c'$) was computed using the following equation:

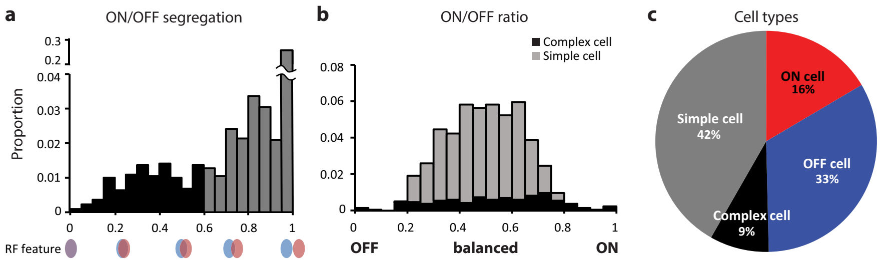
$$c' = \frac{\text{standard deviation}(\delta)}{\text{mean}(\delta)}$$

Uniformity of coverage was computed in two ways: (1) using the original counts of the pixels within the responsive area, and (2) weighting the properties of the pixels by their responsive strength. The coverage for phase was always calculated independently for each orientation and then averaged. Overall, a lower $c'$ value indicates greater uniformity of coverage.

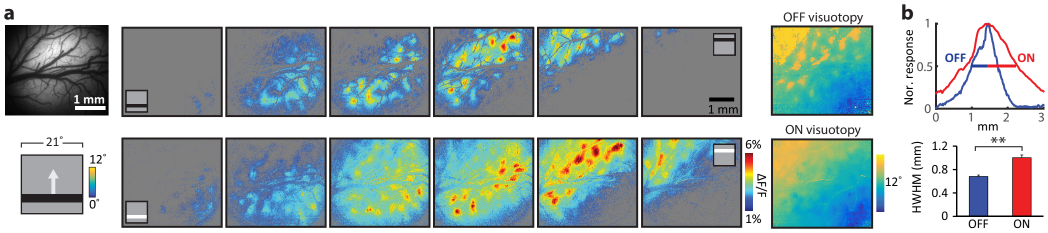
31. Chen, T. W. *et al.* Ultrasensitive fluorescent proteins for imaging neuronal activity. *Nature* **499,** 295–300 (2013).
32. Pologruto, T. A., Sabatini, B. L. & Svoboda, K. ScanImage: flexible software for operating laser scanning microscopes. *Biomed. Eng. Online* **2,** 13 (2003).
33. Sage, D., Prodanov, D., Tinevez, J.Y. & Schindelin, J. MIJ: Making interoperability between ImageJ and Matlab possible. in *ImageJ User & Developer Conference*, 24–26 October 2012, Luxembourg.
34. Peirce, J. W. PsychoPy–Psychophysics software in Python. *J. Neurosci. Methods* **162,** 8–13 (2007).
35. Seelig, J. D. & Jayaraman, V. Feature detection and orientation tuning in the Drosophila central complex. *Nature* **503,** 262–266 (2013).
36. Vogelstein, J. T. *et al.* Fast nonnegative deconvolution for spike train inference from population calcium imaging. *J. Neurophysiol.* **104,** 3691–3704 (2010).
37. Jones, J. P. & Palmer, L. A. An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. *J. Neurophysiol.* **58,** 1233–1258 (1987).
38. Nauhaus, I. *et al.* Orthogonal micro-organization of orientation and spatial frequency in primate primary visual cortex. *Nature Neurosci.* **15,** 1683–1690 (2012).
39. Huang, X. *et al.* Optogenetic assessment of horizontal interactions in primary visual cortex. *J. Neurosci.* **34,** 4976–4990 (2014).
40. Swindale, N. V., Shoham, D., Grinvald, A., Bonhoeffer, T. & Hubener, M. Visual cortex maps are optimized for uniform coverage. *Nature Neurosci.* **3,** 822–826 (2000).



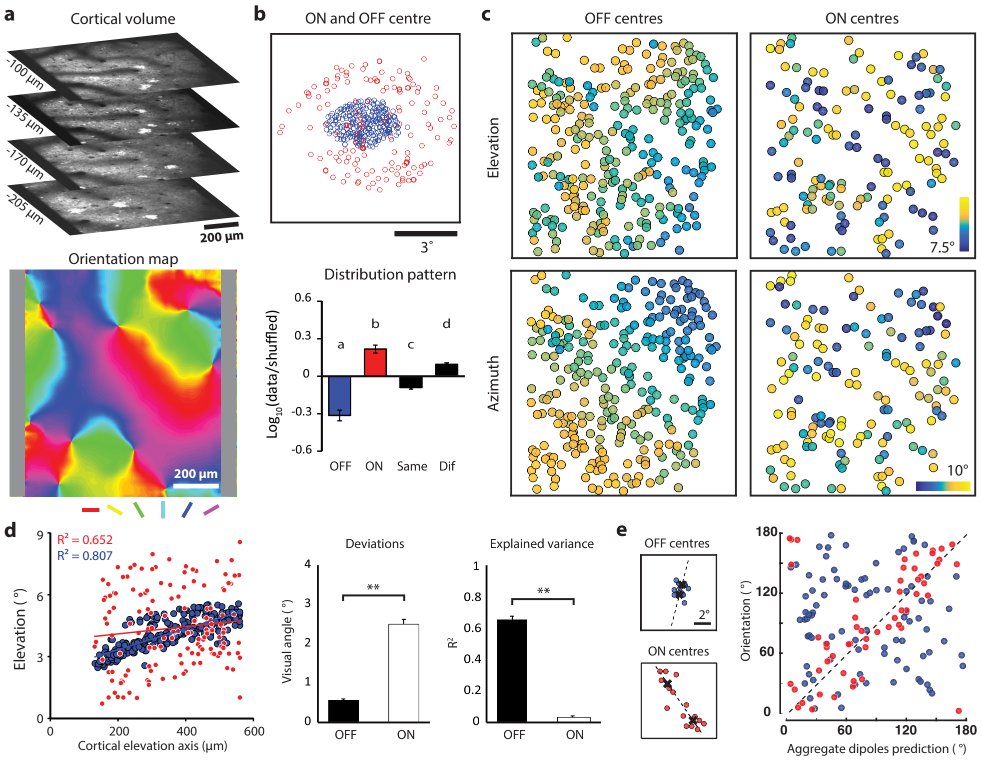

**Extended Data Figure 1 | Robust receptive field estimation from GCaMP6 calcium signal in layer 2/3 neurons. a**, Somatic locations of seven example cells (circles) overlaid on the two-photon field of view. **b**, Raw calcium trace, spatiotemporal receptive field and SNR curves from an example cell in **a**. **c**–**h**, Six different ways to infer the onset time and response strength of neural activity were compared with the original method described in Methods for six of the example cells in **a**. Processed calcium trace before starting inference (blue) and the inferred response (red) are shown on the left. Receptive fields and SNR curves derived from original and alternative methods are shown on the right. **i**–**l**, Change in peak SNR (**i**) and peak time (**j**), receptive field similarity index (**k**) and deviation of the receptive field and subfield centre estimation (**l**) illustrating that the main conclusions regarding receptive field structure and fine visuotopic organization are not altered by the signal processing method employed ($n = 143$ cells from 3 animals). All error bars indicate s.e.m.



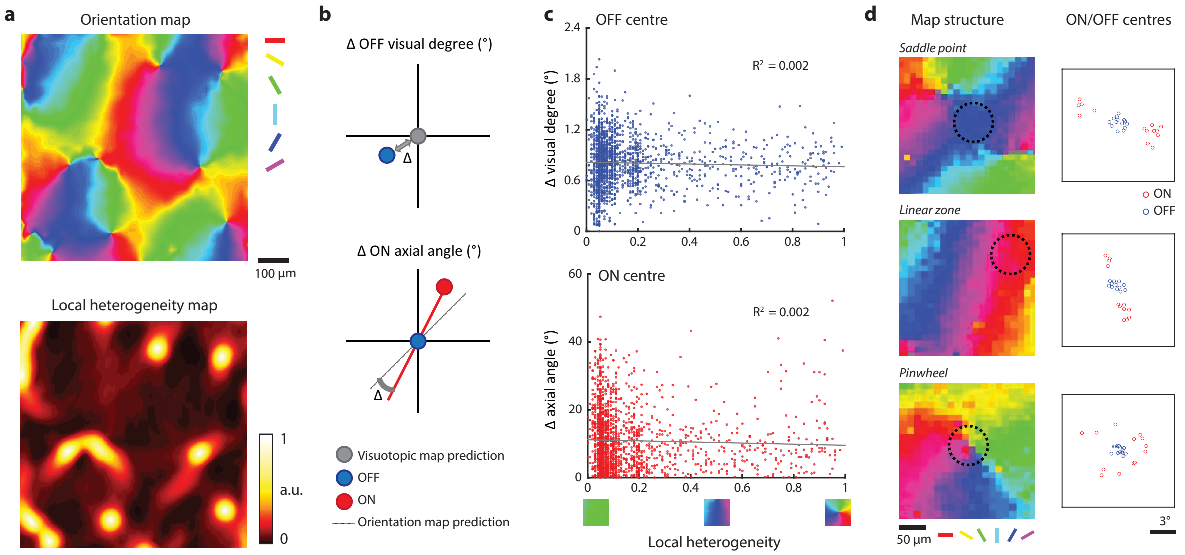

**Extended Data Figure 2 | Cell type categorization in tree shrew primary visual cortex layer 2/3. a**, Distribution of ON/OFF segregation index values for simple and complex cells (see Methods). A value of 0.6 was used to delineate the two classes. **b**, Distribution of ON/OFF ratio values for simple and complex cells. In both **a** and **b**, the proportions are based on the total number of cells; however, the single-sign cell population is not shown in the plots. **c**, Percentage of different classes of neurons in tree shrew visual cortex layer 2/3.



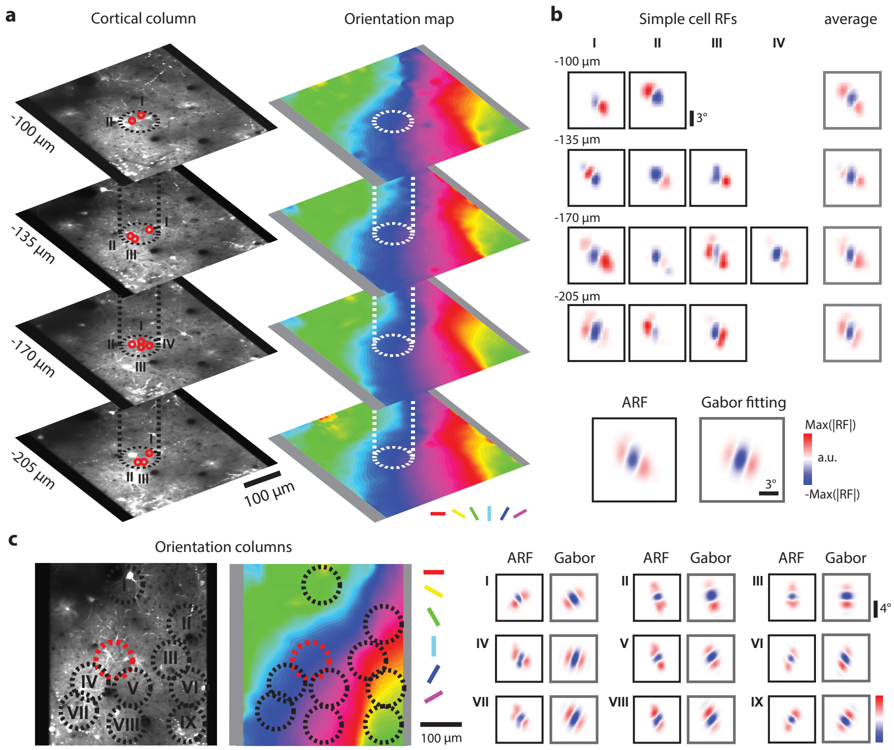

**Extended Data Figure 3 | Cortical spread of light- and dark-evoked activity in epi-fluorescence imaging. a**, Wide-field epi-fluorescence imaging of visual cortex reveals a similar visuotopic progression for the zones of activity found for static light and dark bar stimuli at different locations in elevation. **b**, The bandwidth of the normalized cortical activity pattern, characterized by half width at half maximum (HWHM), shows that light stimuli evoke broader cortical activity patterns than dark stimuli at the same visuotopic location ($n = 21$ stimulus-evoked response maps from 4 animals, $P = 9.6 \times 10^{-5}$, rank-sum test). Error bars indicate s.e.m.



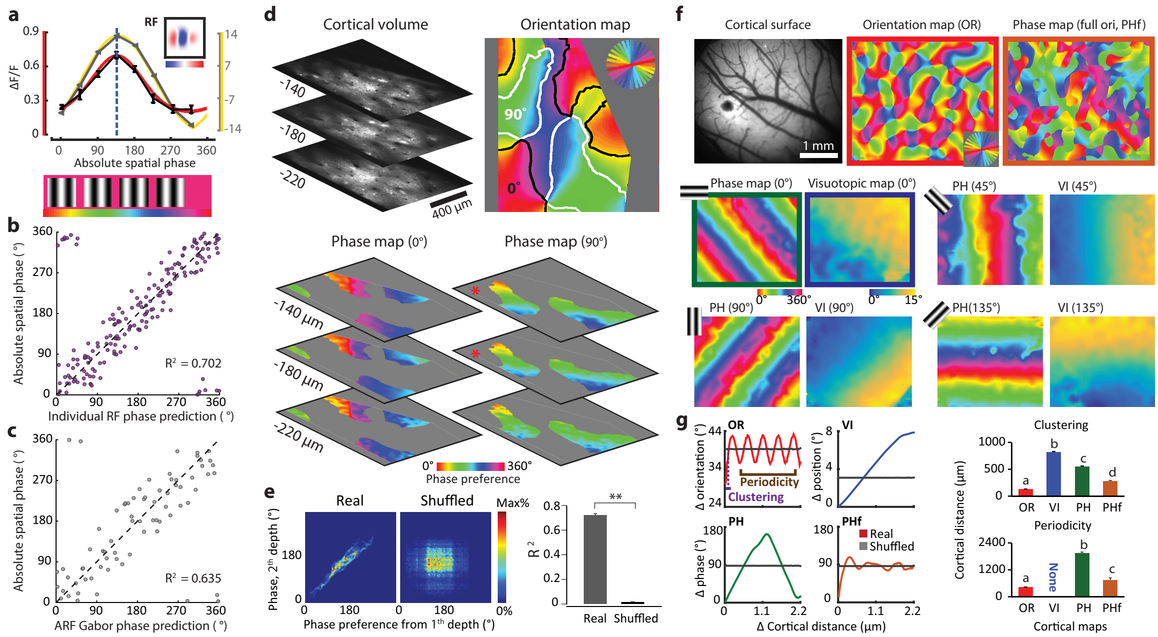
**Extended Data Figure 4 | ON and OFF receptive field organization of single-sign cells. a**, The cortical volume and orientation map of an example imaging area. **b**, The ON and OFF centres from single-sign cells display an arrangement similar to that of simple cells. The bottom plot shows that the distribution pattern of ON and OFF receptive fields is consistent with that of the ON and OFF subfields of simple cells ($n = 8$ imaging areas from 7 animals, Kruskal–Wallis test; compare with Fig. 1c, d; letters indicate groups with statistically significant difference, $P < 0.01$). **c**, The visuotopic organization of ON and OFF receptive field centres was similar to that of simple cell ON and OFF subfields. **d**, The relationship between cortical distance and visuotopic position, demonstrating the difference in visuotopic precision for ON and OFF receptive fields (linear regression). Deviations of the experimental results from the linear fit and explained variance of the smooth visuotopy ($n = 16$ visuotopic maps, combining elevation and azimuth results from 8 imaging areas, $^{**}P < 0.0001$, rank-sum test) are consistent with the results from simple cell ON- and OFF subfields. **e**, Only the displacement of the population ON receptive field centre, but not that of the OFF receptive field centre, can predict the orientation tuning of the orientation column (circular correlation, $n = 68$ cortical columns, $P = 9.51 \times 10^{-3}$ for ON; $n = 89$ cortical columns, $P = 0.586$ for OFF). All error bars indicate s.e.m.



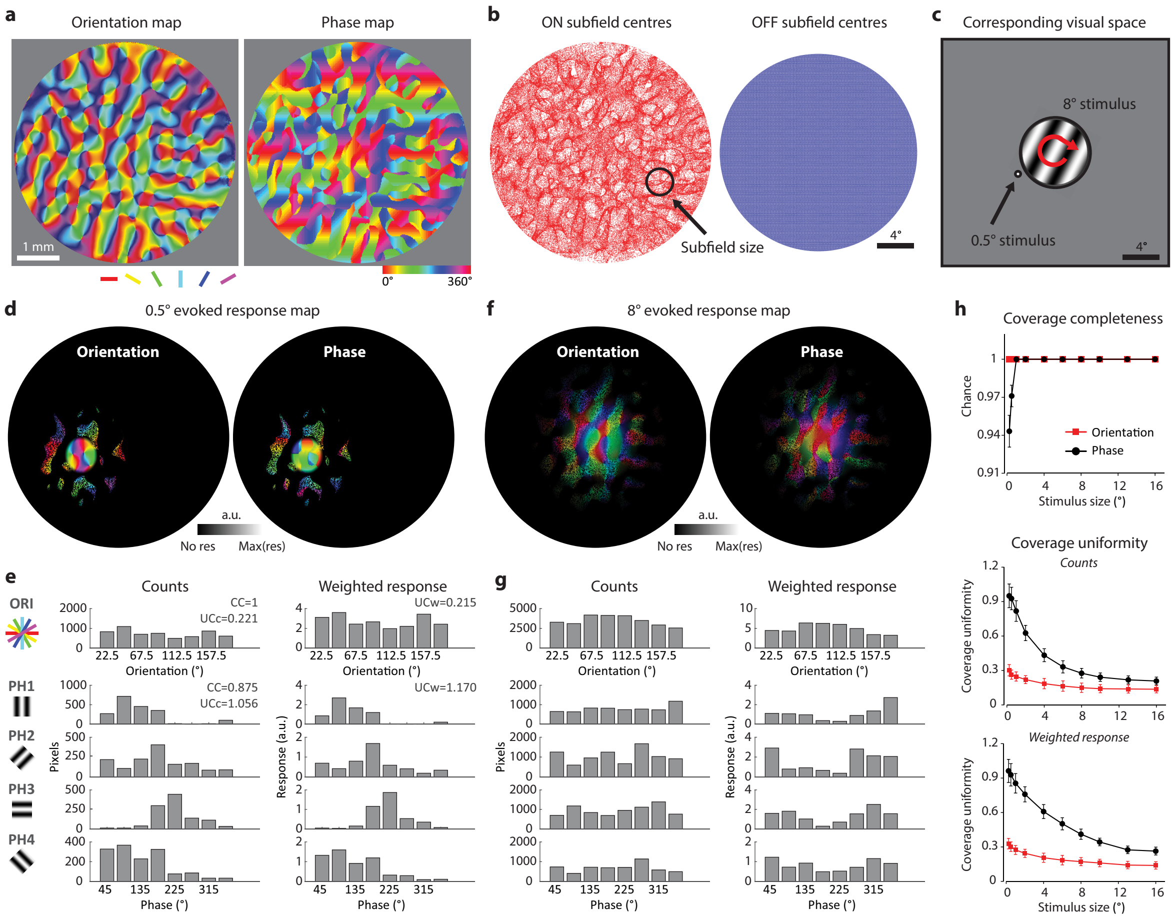

**Extended Data Figure 5 | Visuotopic arrangement of ON and OFF subfields is independent of orientation map structure.** **a**, Example orientation map and local heterogeneity index map. The local heterogeneity index was used to compare ON and OFF subfield arrangement for cortical regions with different orientation map structure. a.u., arbitrary units. **b**, Top, illustration comparing the visuotopic displacement of OFF subfields to the theoretical prediction from a smooth visuotopic map. Bottom, illustration comparing the visuotopic displacement of ON subfields to the orientation map. **c**, Top, visuotopic distortion of OFF subfield centres in relation to the structure of the orientation map. There is no relationship between local heterogeneity and the visuotopic precision of OFF subfields (linear regression, $n = 1{,}811$ cells from 7 animals, $P = 8.2 \times 10^{-2}$). Bottom, axial mismatch of ON subfield centres in relation to the structure of the orientation map. There is no relationship between local heterogeneity and the axial displacement of ON subfield centres (linear regression, $n = 1{,}811$ cells from 7 animals, $P = 9.6 \times 10^{-2}$). **d**, Examples of the ON and OFF subfield centre distributions from 80-µm circular regions (black circles) centred on three distinct regions of the orientation map.




**Extended Data Figure 6 | Contribution of simple cells at different depths to aggregate receptive field of cortical column. a**, An example orientation column at four depths, with two-photon images on the left and the corresponding orientation maps on the right. **b**, Simple cell receptive fields from these four cortical depths. Each RF was normalized by the strongest subfield. The averages of the RFs within each depth appear similar. All the RFs within the orientation column were pooled into an aggregate receptive field (ARF) and then fitted with a 2D Gabor function. **c**, Nine further examples of ARFs from different orientation columns display the same organization: OFF subfield in the centre with ON subfields flanking on two sides.




**Extended Data Figure 7 | Characterizing spatial phase tuning, phase column and phase map. a**, The phase tuning of an example cell (black) and its Gaussian fit (red) compared with the phase tuning curve predicted from its receptive field structure (grey) and its Gaussian fit (yellow). Dashed line depicts the preferred phase derived from the Gaussian fit to the experimental data. **b**, Relationship between absolute phase prediction from receptive field structure and absolute phase tuning measurement ($n = 179$ cells from 2 animals, $P = 1.8 \times 10^{-18}$, circular regression). **c**, Phase preference of the orientation column is well predicted by the phase parameter of the Gabor fit to the ARF ($n = 73$ cortical columns from 5 animals, $P = 1.7 \times 10^{-10}$, circular regression). **d**, Example two-photon phase maps derived from pixel tuning at three cortical depths for both horizontal and vertical orientations. **e**, Comparison of phase preference from different cortical depths (red asterisks in **d**) showing the consistency of columnar structure for spatial phase (rank-sum test for $R^2$ from circular regression, $n = 36$ pairs of maps at different depths from 2 animals, $P = 8.2 \times 10^{-18}$). **f**, Large-scale functional maps visualized by epi-fluorescence imaging. The phase map with full orientation coverage (right) was constructed from four individual phase maps measured independently with four orientations (0°, 45°, 90°, 135°). The phase maps for single orientations with corresponding visuotopic maps are shown separately in the lower two rows. **g**, The statistical structure of functional maps (orientation, phase, visuotopy, and phase with four orientations) summarized by the relationship between the change in cortical distance and the average change in preferred feature (left). Summary comparison of clustering and periodicity of the preferred features of four functional maps from six animals (right). Each map exhibits distinct clustering and periodicity ($n = 32$ sample regions from 6 animals, Kruskal–Wallis test with post-hoc test using Dunn's method, letters indicate groups with statistically significant difference, $P < 0.05$). All error bars indicate s.e.m.




**Extended Data Figure 8 | Simulation based on experimental observations to evaluate completeness and uniformity of coverage for orientation and phase representations. a**, The large-scale orientation preference map derived from intrinsic signal imaging and corresponding phase map predicted from experimental observations (see Methods). **b**, Distribution of ON and OFF subfield centres in visual space predicted from the visuotopic precision and orientation-specific displacement demonstrated in this study. Although the distribution of the ON subfield centres in visual space appears uneven, complete coverage of visual space is achieved when the actual size of the ON subfields is considered (black circle). **c**, Illustration of two of the visual stimuli (8° stimulus in the centre, 0.5° stimulus to the left) used to simulate the evoked response map. **d**, Theoretical stimulus-evoked orientation and phase response maps for sample 0.5° stimulus shown in **c** (see Methods). **e**, Histograms showing the distribution of preferred orientation and phase values for pixels activated in **d**, calculated by counts of the pixels in the responsive region (left) or weighted by the strength of the responses (right). **f**, Theoretical stimulus-evoked orientation and phase response maps for sample 8° stimulus shown in **c** (see Methods). **g**, Histograms showing the distribution of preferred orientation and phase values for pixels activated in **f**, calculated by counts of the pixels in the responsive region (left) or weighted by the strength of the responses (right). **h**, Completeness (top) and uniformity (middle, bottom) of coverage simulated with visual stimuli of various sizes and positions. Complete coverage can be achieved with 1° stimuli, while coverage uniformity continues to improve with increases in stimulus size. The results of spatial phase were always the average results obtained with four different orientations. Error bars indicate s.e.m.

doi:10.1038/nature17939

# Parkinson-associated risk variant in distal enhancer of α-synuclein modulates target gene expression

Frank Soldner[1], Yonatan Stelzer[1], Chikdu S. Shivalila[1,2], Brian J. Abraham[1], Jeanne C. Latourelle[3], M. Inmaculada Barrasa[1], Johanna Goldmann[1], Richard H. Myers[3], Richard A. Young[1,2] & Rudolf Jaenisch[1,2]

**Genome-wide association studies (GWAS) have identified numerous genetic variants associated with complex diseases, but mechanistic insights are impeded by a lack of understanding of how specific risk variants functionally contribute to the underlying pathogenesis[1]. It has been proposed that *cis*-acting effects of non-coding risk variants on gene expression are a major factor for phenotypic variation of complex traits and disease susceptibility. Recent genome-scale epigenetic studies have highlighted the enrichment of GWAS-identified variants in regulatory DNA elements of disease-relevant cell types[2–6]. Furthermore, single nucleotide polymorphism (SNP)-specific changes in transcription factor binding are correlated with heritable alterations in chromatin state and considered a major mediator of sequence-dependent regulation of gene expression[7–10]. Here we describe a novel strategy to functionally dissect the *cis*-acting effect of genetic risk variants in regulatory elements on gene expression by combining genome-wide epigenetic information with clustered regularly-interspaced short palindromic repeats (CRISPR)/Cas9 genome editing in human pluripotent stem cells. By generating a genetically precisely controlled experimental system, we identify a common Parkinson's disease associated risk variant in a non-coding distal enhancer element that regulates the expression of α-synuclein (*SNCA*), a key gene implicated in the pathogenesis of Parkinson's disease. Our data suggest that the transcriptional deregulation of *SNCA* is associated with sequence-dependent binding of the brain-specific transcription factors EMX2 and NKX6-1. This work establishes an experimental paradigm to functionally connect genetic variation with disease-relevant phenotypes.**

Parkinson's disease is the second most common chronic progressive neurodegenerative disorder. The discovery of genes linked to rare Mendelian forms of Parkinson's disease has provided vital clues to the molecular and cellular pathogenesis of the disease[11]. However, over 90% of Parkinson's disease cases do not show Mendelian inheritance patterns, suggesting that sporadic, late-onset Parkinson's disease results from a complex interaction between genetic and environmental risk factors. While coding mutations and genomic multiplications of the *SNCA* gene cause familiar Parkinson's disease, GWAS have identified *SNCA* as one of the strongest risk loci associated with the sporadic form of the disease, suggesting a pivotal role in the pathogenesis of Parkinson's disease[12]. Genomic duplications of *SNCA* indicate that an increase by 50% in *SNCA* expression is sufficient to develop an autosomal-dominant form of the disease, suggesting that Parkinson's disease associated risk variants might lead to a subtle increase in *SNCA* expression[13–15]. To analyse such small changes in gene expression despite considerable technical and biological heterogeneity of *in vitro* human pluripotent stem cell culture and differentiation systems, we conceived a novel experimental approach which allowed us to reliably quantify the consequences of targeted genetic modifications on transcription by analysing the *cis*-acting effects on allele-specific expression. Figure 1a, b illustrates how the heterozygous deletion or exchange of a candidate regulatory element through *cis*-regulatory effects on expression is predicted to modulate allele-specific gene expression when measured as the ratio between the modified and the non-targeted allele.

To analyse precisely the expression of two individual alleles in a single multiplex reaction, we adapted TaqMan SNP genotyping assays to quantitative reverse transcription polymerase chain reaction (qRT–PCR; Extended Data Fig. 1a). A common SNP (rs356165 A/G, referred to as *SNCA* 'reporter SNP') was identified in the 3′ UTR of *SNCA* in two human embryonic stem (ES) cell lines and a common primer pair and allele-specific TaqMan probes conjugated with different fluorophores were used to distinguish between the two alleles (FAM to detect the A allele and VIC to detect the G allele). To validate this approach, we simulated allele-biased samples over a wide range of *SNCA* expression ratios by mixing cDNAs from two types of human induced pluripotent stem cell (iPS cell) derived neurons[16] that are homozygous for either the A or the G allele at the reporter SNP. Multiplex allele-specific qRT–PCR analysis robustly quantified the expression of each individual allele in the mixed samples (Fig. 1c and Extended Data Fig. 1b) with the relative allele-specific expression of the two alleles closely correlating with the expected ratio (Fig. 1d). Comparing neurons derived from isogenic cultures in parallel at different time points during terminal differentiation (Extended Data Fig. 2) revealed considerable differences in total *SNCA* expression (Fig. 1e). In contrast, allele-specific expression remained constant across all conditions (Fig. 1f). These data indicate that allele-specific TaqMan qRT–PCR analysis robustly allows detection of small effects on allele-specific expression, independent of cellular heterogeneity due to *in vitro* differentiation and maturation.

A recent analysis in post-mortem tissue from adult brain identified a significant enrichment of Parkinson's disease associated SNPs within distal enhancers[17], consistent with the notion that GWAS variants in regulatory elements can be used to prioritize functional disease-relevant risk alleles[18,19]. To identify candidate risk variants in enhancers, we intersected Parkinson's disease associated SNPs in the *SNCA* locus (463 SNPs, $P < 5 \times 10^{-8}$ provided by PDGene database)[12] with publicly available epigenetic data (NIH Roadmap Epigenomics Consortium; http://www.roadmapepigenomics.org)[20]. Ranking of all of Parkinson's disease associated SNPs in the *SNCA* locus based on cumulative overlap with enhancer-associated marks[3,21,22] such as H3K4me1, H3K27ac and DNase I hypersensitive sites (DHSs), revealed that the top seven risk variants were localized to two distal enhancer elements (intron-4 enhancer and 3′ UTR enhancer, Fig. 2a, Extended Data Fig. 3a–c and Supplementary Table 1) with both displaying an active epigenetic signature in the substantia nigra and in human ES-cell-derived neurons (Extended Data Fig. 3b, d). Because SNP-specific changes are thought to modify enhancer activity by altering transcription factor (TF) binding[7–10], we analysed predicted TF binding by scanning for known binding sequence motifs comparing both alternative genotypes for each Parkinson's disease associated SNP. This analysis indicated that the risk

[1]The Whitehead Institute, 9 Cambridge Center, Cambridge, Massachusetts 02142, USA. [2]Department of Biology, Massachusetts Institute of Technology, 31 Ames Street, Cambridge, Massachusetts 02139, USA. [3]Department of Neurology, Boston University School of Medicine, Boston, Masssachusetts 02118, USA.

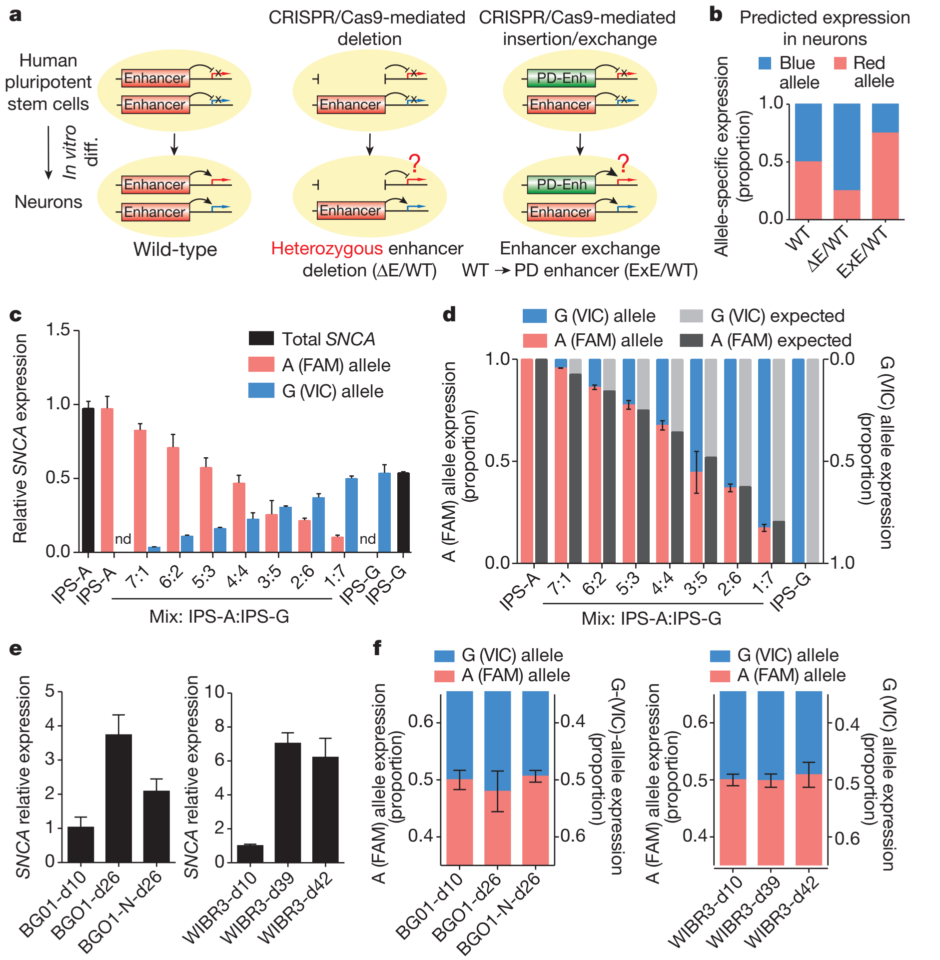



**Figure 1 | Strategy to analyse *cis*-regulatory effects of genetic variants on allele-specific expression of *SNCA*.** **a**, Schematic illustration of the experimental strategy to study the function of Parkinson's disease associated regulatory elements in human pluripotent-stem-cell-derived neurons by CRISPR/Cas9-mediated heterozygous deletion (ΔE) or exchange/insertion (ExE) of risk-associated enhancer sequences. PD, Parkinson's disease. **b**, Expected effect on relative allele-specific gene expression in human pluripotent-stem-cell-derived neurons resulting from enhancer modifications described in **a**. **c**, Allele-specific *SNCA* expression analysis of allele-biased samples with indicated allele ratios. Allele-biased samples were generated by mixing human iPS-cell-derived neurons homozygous for either the A (IPS-A) or G allele (IPS-G) at rs356165. The relative expression of each allele was normalized to the total expression of *SNCA* (black bars) in the respective sample. **d**, qRT–PCR analysis for relative expression of the A (FAM) allele (red bars) and the G (VIC) allele (blue bars) in allele-biased samples described in **c**, (calculated as proportion) compared with expected relative proportions (light and dark grey bars). **e**, Analysis of total *SNCA* expression at different time points during *in vitro* differentiation of human ES-cell-derived neurons (BGO1 and WIBR3). **f**, Allele-specific *SNCA* expression of human ES-cell-derived neurons described in **e**, normalized to differentiation day-10 cultures. Shown are mean values ± s.d. ($n = 3$). nd, not detected. Source Data for this figure are available online.

variant rs356168 in the intron-4 enhancer is a TF binding hotspot with the highest number of predicted genotype-dependent differential binding of all of Parkinson's disease associated SNPs in the *SNCA* locus (Extended Data Fig. 3b, c; Supplementary Table 1).

To analyse the function of the intron-4 enhancer element, we deleted 500 bps of the epigenetically marked enhancer region containing the Parkinson's disease associated risk SNPs rs356168 and rs3756054 by CRISPR/Cas9-mediated genome editing (Fig. 2b and Extended Data Figs 3b and 4a). We subsequently reinserted an allelic series of intron-4 enhancer elements harbouring all possible genotype combinations for rs356168 and rs3756054 into the enhancer-deleted cells (Fig. 2b and Extended Data Fig. 4a) to dissect sequence-specific effects of each risk variant. Southern blot analysis and genomic sequencing confirmed correct integration of the targeted enhancer elements in *cis* with the A (FAM) allele reporter SNP (Fig. 2b and Extended Data Fig. 4). To analyse the effect of each enhancer genotype on *SNCA* expression, we differentiated between 2 and 4 individual clones targeted for each enhancer element into neural precursors or mixed neuronal cultures. Initial expression analysis for total *SNCA* and markers for neuronal and astrocytic differentiation showed no consistent differences between the genotypes (Extended Data Fig. 5a–c). In contrast, allele-specific qRT–PCR analysis revealed that neural precursors and neurons carrying the G allele at rs356168 showed a highly significant increase in expression of the A (FAM) reporter SNP compared with cells carrying the A allele at rs356168 and the homozygous enhancer deleted controls (Fig. 2c, d and Extended Data Fig. 5d–g). Because the inserted enhancer elements are in *cis* with the A (FAM) reporter SNP (Fig. 2b), we conclude that only insertion of enhancer sequences carrying the G allele at rs356168 results in increased expression of *SNCA*, whereas the A allele at this SNP has no effect compared with enhancer deleted controls. This effect was independent of the adjacent risk variant rs3756054, which has no considerable effect on allele-specific *SNCA* expression (Fig. 2c, d and Extended Data Fig. 5d–g). Genotyping of the parental human ES cell line revealed that WIBR3 is heterozygous at SNP rs356168 with the A allele in *cis* with the A (FAM) reporter SNP and the G allele in *cis* with the G (VIC) reporter SNP (Extended Data Fig. 5h). The homozygous deletion of the intron-4 enhancer resulted in decreased allele-specific expression of the G (VIC) allele (Extended Data Fig. 5i) consistent with the observation that only the G allele at rs356168 significantly modifies the expression of the *cis*-regulated allele. Our results suggest that the intron-4 enhancer element regulates the transcription of *SNCA* in human neural precursor cells and neurons, and that the common SNP at rs356168 represents a functional risk variant, with the G allele causing increased expression. This is consistent with the GWAS data, which identify the non-active A allele at rs356168 as a protective allele with an odds ratio (OR) of 0.79 (0.76–0.81)[12]. In contrast, the Parkinson's disease associated SNP rs3756054 has no effect on *SNCA* expression suggesting that this variant is in linkage disequilibrium (LD) with the risk-modifying SNP.

To further support a functional effect of rs356168, we performed an expression quantitative trait loci (eQTL) analysis of total *SNCA* mRNA levels measured by qRT–PCR in 127 post-mortem frontal cortex samples (86 Parkinson's disease samples, 41 control samples). A significant increase in total *SNCA* levels ($P = 0.031$; linear regression analysis) was observed in carriers of the rs356168 risk allele (Extended Data Fig. 6b). In comparison, only a modest and not significant increase ($P = 0.33$) in expression was observed for rs356229 (as proxy for top reported GWAS SNP[12] rs356182; $R^2 = 0.62$), indicating that rs356168 more precisely predicts *SNCA* expression levels.

To assess the extent to which rs356168 could explain the associations observed in the *SNCA* region in Parkinson's disease GWAS[12], we completed a baseline and conditional analysis in 5 publicly available Parkinson's disease GWAS cohorts totalling 6,014 cases and 9,119

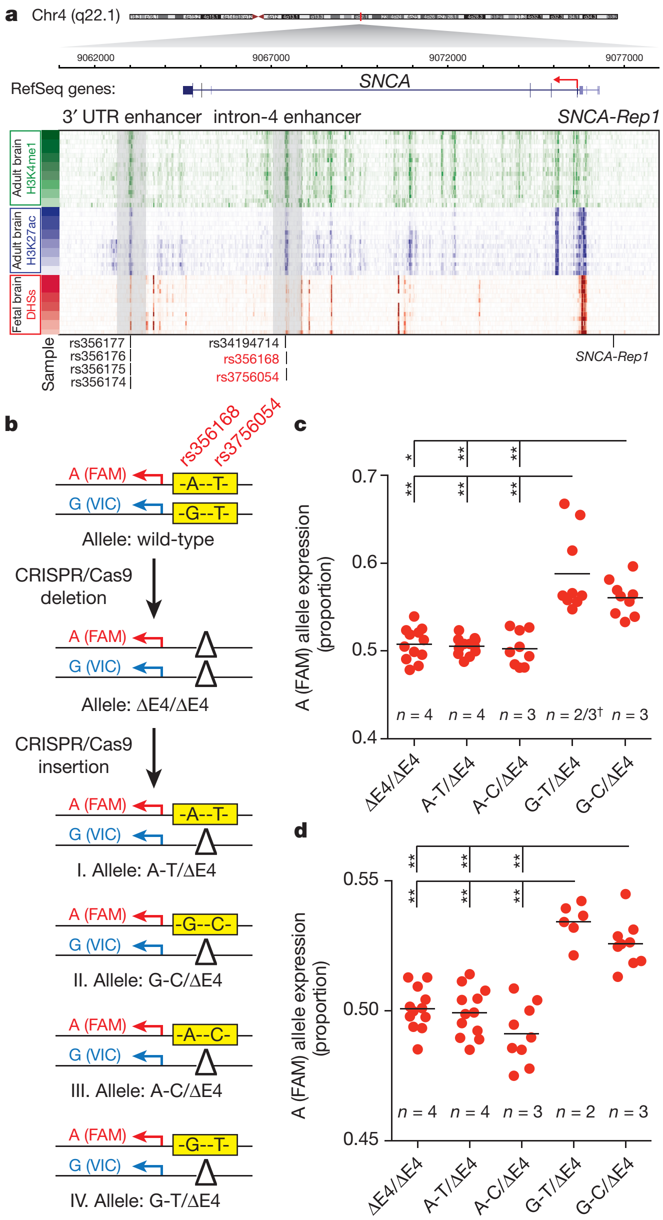


**Figure 2 | Identification of a functional *cis*-acting Parkinson's disease associated SNP in an intronic enhancer element of *SNCA*. a**, Heatmap of H3K4me1 and H3K27ac ChIP-seq and DHSs-enrichment tracks for several brain regions in the *SNCA* locus (for details see Extended Data Fig. 3a; data provided by NIH Roadmap Epigenomics Consortium; http://www.roadmapepigenomics.org). Shown are the locations of *SNCA-Rep1* and Parkinson's disease associated SNPs overlapping with two proximal enhancer elements (3′ UTR enhancer and intron-4 enhancer) highlighted by light grey boxes. **b**, Schematic illustration of the CRISPR/Cas9-mediated strategy to delete and subsequently insert intron-4 enhancer elements with indicated Parkinson's disease associated risk SNP genotypes at rs356168 and rs3756054. Targeted clones carrying inserted risk alleles in *cis* with the A (FAM) reporter SNP (confirmed by genomic sequencing-based phase-reconstruction) were used for subsequent analysis described in **c**, **d**. **c**, **d**, Relative allele-specific *SNCA* expression in neural precursors (**c**) and mixed neuronal cultures (**d**) (differentiation day 25) derived from targeted cell lines with indicated intron-4 enhancer alleles compared to human ES cells carrying homozygous enhancer deletions (ΔE4/ΔE4) (expression was normalized relative to ΔE4/ΔE4 cell lines). Data are presented as dot plot; each dot represents mean of 3 technical replicates. Allele-specific expression for each clone was analysed in 3 independent biological replicate experiments and combined according to genotypes. Black lines indicate mean expression for each genotype; *n* indicates number of independently targeted clones per genotype, † indicates an additional sub-clone derived from one of the two targeted clones for this genotype. Statistical differences between genotypes were calculated using one-way ANOVA (alpha = 0.05) followed by Tukey's multiple comparison test based on allele-specific expression of all biological replicates. $*P < 0.001$; $**P < 0.0001$; Source Data and detailed statistical analysis are provided online.

controls. The effect of the top reported GWAS SNP rs356182 was reduced 28% from an OR of 1.32 to 1.23 and the statistical significance was attenuated from a genome-wide significant $P$ value of $1.1 \times 10^{-21}$ to $3.0 \times 10^{-6}$ (Extended Data Fig. 6a). As expected, the independent 5′ region SNP rs7681154 which was revealed in the initial GWAS conditional analysis[12] continued to show a significant independent effect when conditioning on rs356168.

Although this type of analysis cannot identify by itself functional risk variants, the results are consistent with the hypothesis that multiple functional variants with small size effects contribute to the overall association and heritability of the *SNCA* locus with sporadic Parkinson's disease (common disease-common variant hypothesis). The *in vitro* observed changes in allele-specific expression translate roughly to an increase of total *SNCA* expression of 1.06 times in neurons and 1.18 times in neural precursors. Given that a 1.5-fold increase in *SNCA* expression is sufficient to cause a familial autosomal-dominant form of Parkinson's disease, our data support the notion that a modest life-long increase of *SNCA* expression may represent the molecular basis of an increased risk to develop Parkinson's disease for carriers of the G allele.

Sequence-specific changes in chromatin state associated with differential enhancer activities, have been proposed as a mechanism for SNP-dependent *cis*-regulatory effects on gene expression[7–10]. Chromatin immunoprecipitation (ChIP) followed by qRT–PCR (ChIP-qRT–PCR) for H3K4me1 and H3K27ac and ChIP-sequence (ChIP-seq) analysis for H3K27ac in neurons carrying all distinct genotypes indicated, that all intron-4 enhancer elements display a chromatin signature for an active enhancer (Extended Data Fig. 7). Normalization of intron-4 enhancer ChIP-seq reads relative to the 3′ UTR enhancer of *SNCA* suggest a slight trend of increased H3K27ac read density in neurons carrying the G allele at rs356168 (Extended Data Fig. 7e) consistent with sequence-dependent changes of chromatin state and enhancer activity. However, considering the small SNP-dependent effect on transcription, it is difficult to distinguish between sequence-specific changes and technical variations resulting from *in vitro* differentiation associated variability.

Differential TF binding is considered a major mediator of sequence-specific effects of distal enhancers on gene regulation[7–10]. We compiled a list of all TFs predicted to show SNP-specific binding at rs356168 by scanning the sequences of both alternative alleles for known binding motifs (Supplementary Table 2). We selected 10 candidate TFs (see Methods) and performed ChIP-qRT–PCR to identify candidates that specifically bind to the intron-4 enhancer element in *in vitro* differentiated neurons. This analysis identified binding of the brain-expressed TFs NKX6-1 and EMX2 (Fig. 3a and Extended Data Fig. 8a). Immunostaining confirmed the expression of EMX2 and NKX6-1 in human ES-cell-derived neurons and, to a lesser extent, in interspersed astrocytes (Extended Data Fig. 9f–j). Single-molecule mRNA fluorescence *in situ* hybridization (FISH) showed that more than 40% of the cultured cells were positive for one of the two TFs, while only a smaller fraction (around 20%) expressed both factors simultaneously (Extended Data Fig. 9a–e). Importantly, electrophoretic mobility shift assay analysis (EMSA) revealed a clear SNP-dependent binding of EMX2 and NKX6-1 with preference for the protective lower *SNCA* expressing A allele at rs356168 (Fig. 3b, c and Extended Data Fig. 8b–e). These results suggest a model in which the sequence-specific binding of EMX2 and NKX6-1 at a distal enhancer element represses enhancer activity and thus modulates *SNCA* expression (Fig. 4). These data further suggest that the same enhancer elements may be regulated in distinct

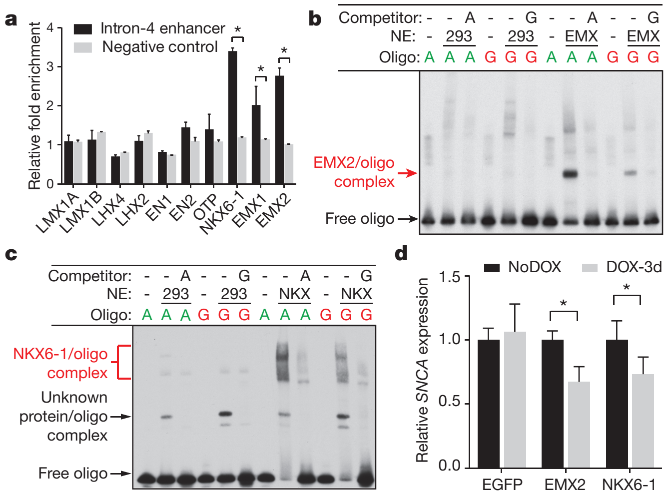



**Figure 3 | Sequence-specific effect of Parkinson's disease associated risk variants on binding of brain-expressed TFs EMX2 and NKX6-1 at the *SNCA* intron-4 enhancer. a**, ChIP-qRT–PCR analysis for binding of indicated TFs at the intron-4 enhancer element compared with a negative control region in the *SNCA* locus (calculated as fold enrichment compared with IgG isotype control, shown are mean values ± s.d., $n = 2$). Statistical significance was determined using a *t*-test followed by the Holm–Šídák method to correct for multiple comparisons (alpha = 0.05). $*P < 0.0001$. **b**, **c**, EMSA analysis for SNP-genotype-specific binding of EMX2 (**b**) or NKX6-1 (**c**) to oligonucleotides (oligo) harbouring the indicated genotype at rs356168 (A/G allele). Binding was analysed in nuclear extracts (NE) from wild-type (293) or EMX2 (**b**, EMX) or NKX6-1 (**c**, NKX) overexpressing HEK293 cells. Red arrows point to oligonucleotide-specific binding which is lost in the presence of unlabelled competitor oligonucleotides (with indicated genotype at rs356168; 200X). **d**, *SNCA* expression analysis following doxycycline-induced overexpression of EGFP, EMX2 or NKX6-1 for 3 days in terminally differentiated neurons (differentiation day 21). Shown are mean values ± s.d. ($n = 10$) of relative *SNCA* expression in doxycycline-induced cells (DOX-3d) compared with the corresponding untreated controls (NoDOX). Results are representative of two different experiments. Statistical significance was calculated using a *t*-test followed by the Holm–Šídák method to correct for multiple comparisons (alpha = 0.05). $*P < 0.0001$. Source Data for this figure are available online.

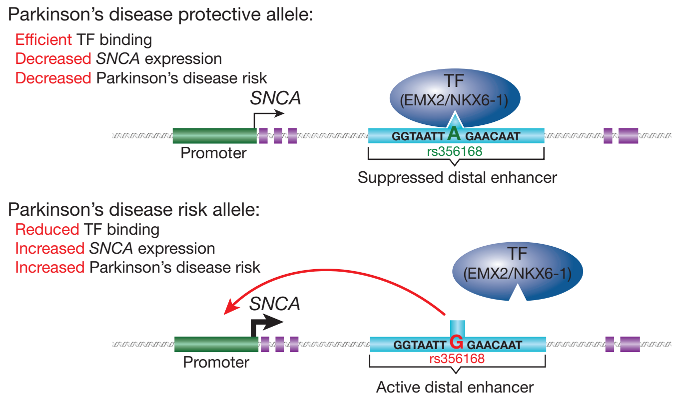



**Figure 4 | Proposed model describing the correlation between SNP-dependent TF binding, *SNCA* expression and Parkinson's disease risk.** Carriers of the A allele at rs356168 (Parkinson's disease protective allele) show efficient binding of the brain-specific TFs EMX2 and NKX6-1 at the distal intron-4 enhancer, which results in a suppressed distal enhancer and consequently lower expression of *SNCA* associated with a reduced risk to develop Parkinson's disease. In contrast, carriers of the G allele at rs356168 (Parkinson's disease risk allele) show reduced TF binding, which results in an active distal enhancer leading to increased expression of *SNCA* and increased risk of developing Parkinson's disease.

neuronal populations by different TFs, as it does not seem crucial that both proteins are expressed in the same cells. Indeed, the overexpression of EMX2 and NKX6-1 in terminally differentiated neurons, using a highly controlled doxycycline inducible system, resulted in significant downregulation of *SNCA* expression (Fig. 3d) further supporting this hypothesis. Our results are consistent with mouse models demonstrating a similar mechanism as repressors of enhancer function[23–25].

Several genetic association studies identified a polymorphic microsatellite repeat region (*SNCA-Rep1*) 10 kb upstream of the transcription start site (Fig. 2a) associated with Parkinson's disease risk. These studies suggest that individuals homozygous for a shorter repeat region (*Rep1-257* or *Rep1-259*) have a significant lower risk of developing Parkinson's disease than individuals carrying the longer forms (*Rep1-261* or *Rep1-263*)[26]. Functional studies suggested an enhancer-like function based on the *cis*-regulatory correlation between the *SNCA-Rep1* repeat length and *SNCA* expression[27,28]. To test whether *SNCA-Rep1* length influences *SNCA* expression in human neurons, we deleted the entire repeat region and subsequently inserted representative alleles for each of the 4 reported repeat length alleles[26] (Extended Data Fig. 10a). Genomic sequencing, Southern blot analysis and fragment length analysis confirmed correct integration of the respective alleles (Extended Data Fig. 10b–d). Multiple sub-clones of each of the *SNCA-Rep1* alleles were differentiated into neurons and analysed for allele-specific *SNCA* expression. Athough individual clones varied considerably, no repeat length dependent effect of *SNCA-Rep1* on *SNCA* expression was detected in two human ES cell lines (Extended Data Fig. 10e, f). Moreover, the deletion of the entire repeat region thought to have the strongest effect an *SNCA* expression[27] did not significantly alter the allele-specific expression compared with the parental human ES cell lines or any of the other *SNCA-Rep1* alleles. This result conflicts with the microsatellite repeat region exerting a *cis*-regulatory effect on of *SNCA* expression. It is possible that difficulties in controlling the experimental variables in neuroblastoma cells[28] or transgenic mice[27] affected the validity of the previous conclusions. As *in vitro* differentiated cells allow only for the analysis of early events, we cannot completely exclude an effect of the *SNCA-Rep1* element at later time points or in combination with additional factors such as environmental stress.

The generation of patient-derived human iPS cells, which carry all pathogenic genetic alterations, is attractive for the study of diseases. However, significant biological heterogeneity due to differences in genetic background, variation in human iPS cell isolation and *in vitro* differentiation present a serious limitation for identifying a disease-relevant phenotype in the culture dish[29]. This is particularly relevant for sporadic diseases likely displaying only subtle *in vitro* phenotypes. Here we describe an alternative experimental approach to identify functional risk variants based on three recent innovations in genetics and molecular biology: (i) the prioritization of GWAS-identified risk variants in regulatory elements such as distal enhancers annotated based on genome-scale epigenetic data; (ii) the generation of genetically controlled isogenic pluripotent stem cell lines in which specific disease-associated genetic variants are the sole modified experimental variable using efficient gene-editing technologies such as the CRISPR/Cas9 system[30]; and (iii) the analysis of *cis*-acting effects of candidate variants on allele-specific gene expression through deletion or exchange of disease-associated regulatory elements. This approach eliminates the effect of system inherent variability such as *in vitro* differentiation and results in an internally controlled experimental system, which allows robust and reproducible identification of *cis*-acting sequence-specific effects on gene regulation. Importantly, the experimental paradigm established here is not only relevant for Parkinson's disease, but is generally applicable for mechanistic studies of the molecular consequences of risk alleles associated with other diseases.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 17 July 2015; accepted 24 March 2016.**
**Published online 20 April 2016.**

1. McClellan, J. & King, M.-C. Genetic heterogeneity in human disease. *Cell* **141,** 210–217 (2010).
2. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473,** 43–49 (2011).
3. Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337,** 1190–1195 (2012).
4. Trynka, G. *et al.* Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nature Genet.* **45,** 124–130 (2013).
5. Hnisz, D. *et al.* Super-enhancers in the control of cell identity and disease. *Cell* **155,** 934–947 (2013).
6. Degner, J. F. *et al.* DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* **482,** 390–394 (2012).
7. Kilpinen, H. *et al.* Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. *Science* (2013).
8. McVicker, G. *et al.* Identification of genetic variants that affect histone modifications in human cells. *Science* (2013).
9. Kasowski, M. *et al.* Extensive variation in chromatin states across humans. *Science* (2013).
10. Leung, D. *et al.* Integrative analysis of haplotype-resolved epigenomes across human tissues. *Nature* **518,** 350–354 (2015).
11. Singleton, A. B., Farrer, M. J. & Bonifati, V. The genetics of Parkinson's disease: progress and therapeutic implications. *Mov. Disord.* **28,** 14–23 (2013).
12. Nalls, M. A. *et al.* Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson's disease. *Nature Genet.* **46,** 989–993 (2014).
13. Devine, M. J., Gwinn, K., Singleton, A. & Hardy, J. Parkinson's disease and α-synuclein expression. *Mov. Disord.* **26,** 2160–2168 (2011).
14. Miller, D. W., Hague, S. M., Clarimon, J. & Baptista, M. α-synuclein in blood and brain from familial Parkinson disease with SNCA locus triplication. *Neurology* **62,** 1835–1838 (2004).
15. Kim, H. J., Jeon, B. S., Yoon, M. Y. & Park, S. S. Increased expression of alpha-synuclein by SNCA duplication is associated with resistance to toxic stimuli. *J. Mol. Neurosci.* **47,** 249–255 (2012).
16. Soldner, F. *et al.* Parkinson's disease patient-derived induced pluripotent stem cells free of viral reprogramming factors. *Cell* **136,** 964–977 (2009).
17. Vermunt, M. W. *et al.* Large-scale identification of coregulated enhancer networks in the adult human brain. *Cell Rep.* **9,** 767–779 (2014).
18. Ward, L. D. & Kellis, M. Interpreting noncoding genetic variation in complex traits and human disease. *Nature Biotechnol.* **30,** 1095–1106 (2012).
19. Rivera, C. M. & Ren, B. Mapping human epigenomes. *Cell* **155,** 39–55 (2013).
20. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518,** 317–330 (2015).
21. Creyghton, M. P. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl Acad. Sci. USA* **107,** 21931–21936 (2010).
22. Rada-Iglesias, A. *et al.* A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470,** 279–283 (2011).
23. Mariani, J. *et al.* Emx2 is a dose-dependent negative regulator of Sox2 telencephalic enhancers. *Nucleic Acids Res.* **40,** 6461–6476 (2012).
24. Ligon, K. L. Loss of Emx2 function leads to ectopic expression of Wnt1 in the developing telencephalon and cortical dysplasia. *Development* **130,** 2275–2287 (2003).
25. Schaffer, A. E., Freude, K. K., Nelson, S. B. & Sander, M. Nkx6 transcription factors and Ptf1a function as antagonistic lineage determinants in multipotent pancreatic progenitors. *Dev. Cell* **18,** 1022–1029 (2010).
26. Farrer, M. *et al.* α-synuclein gene haplotypes are associated with Parkinson's disease. *Hum. Mol. Genet.* **10,** 1847–1851 (2001).
27. Cronin, K. D. *et al.* Expansion of the Parkinson disease-associated *SNCA*-Rep1 allele upregulates human α-synuclein in transgenic mouse brain. *Hum. Mol. Genet.* **18,** 3274–3285 (2009).
28. Chiba-Falek, O., Kowalak, J. A., Smulson, M. E. & Nussbaum, R. L. Regulation of alpha-synuclein expression by poly (ADP ribose) polymerase-1 (PARP-1) binding to the NACP-Rep1 polymorphic site upstream of the SNCA gene. *Am. J. Hum. Genet.* **76,** 478–492 (2005).
29. Soldner, F. & Jaenisch, R. Medicine. iPSC disease modeling. *Science* **338,** 1155–1156 (2012).
30. Soldner, F. *et al.* Generation of isogenic pluripotent stem cells differing exclusively at two early onset Parkinson point mutations. *Cell* **146,** 318–331 (2011).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** We acknowledge the investigators of the original Parkinson's disease GWAS meta-analysis, PDGene and Menno Creyghton at the Hubrecht Institute for sharing data used for this study. We would like to thank B. Zaidi for helpful discussion in conceiving this work. We thank R. Alagappan, D. Fu and T. Lungjangwa for technical support and help in human ES cell culture and molecular biology and A. D'Alessio for helpful advice for performing ChIP-seq experiments. We would like to thank P. Wisniewski, M. Ly, C. Zollo and C. Araneo of the Whitehead Institute FACS facility for their help with cell sorting, T. Volkert, J. Love and S. Gupta at the Whitehead Genome Technologies Core for Solexa sequencing, W. Salmon and N. Watson from the W. M. Keck Biological Imaging Facility and T. DiCesare for help with the figure illustrations. We thank all the members of the Jaenisch laboratory for discussions and comments on the manuscript. R.J. was supported by NIH grants 1R01NS088538-01 and 2R01MH104610-15 and by Qatar National Research Fund grant number NPRP 5-531-1-094.

**Author Contributions** F.S. and R.J. conceived the project, designed and supervised the experiments, interpreted results and wrote the paper with input from all authors. Y.S. assisted with ChIP experiments and contributed to the preparation of the manuscript. C.S.S. assisted with CRISPR/Cas9 genome editing. B.J.A. performed computational analysis of epigenetic datasets and prioritization of Parkinson's disease associated SNPs. J.C.L. performed conditional GWAS and eQTL analysis. M.I.B. performed computational analysis to predict TF binding. J.G. performed single-molecule mRNA FISH analysis. R.H.M. and R.A.Y. supervised data analysis, interpreted results and contributed to the manuscript preparation. F.S. performed all other experiments.

**Author Information** Raw data from ChIP-seq analysis have been submitted to the NCBI Gene Expression Omnibus under the accession number GSE71278. Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to R.J. (jaenisch@wi.mit.edu).

## METHODS

**Human ES cell and human iPS cell culture.** Human ES cell and human iPS cell culture conditions have been described previously[30]. Human iPS cell[16] and human ES cell lines WIBR3 (Whitehead Institute Center for Human Stem Cell Research, Cambridge, MA)[31] and BGO1 (NIH Code: BG01; BresaGen, Athens, GA) were maintained on mitomycin C-inactivated mouse embryonic fibroblast (MEF) feeder layers in human ES cell medium (DMEM/F12 (Invitrogen) supplemented with 15% fetal bovine serum (FBS) (Hyclone), 5% KnockOut Serum Replacement (Invitrogen), 1 mM glutamine (Invitrogen), 1% nonessential amino acids (Invitrogen), 0.1 mM β-mercaptoethanol (Sigma) and 4 ng $ml^{-1}$ FGF2 (R&D systems)). Cultures were passaged every 5 to 7 days either manually or enzymatically with collagenase type IV (Invitrogen; 1.5 mg $ml^{-1}$). All experiments in this study were performed in a sub-clone of WIBR3 and BGO1 with a targeted inframe insertion of *EGFP* into the *NURR1* locus (data not shown) which should not influence the results reported here and will be described in detail in a separate publication. The identities of all parental human ES and human iPS cell lines were confirmed by DNA fingerprinting and all cell lines were regularly tested to exclude mycoplasma contaminations using a PCR based assay.

**NPC culture and terminal differentiation.** Differentiation into neural precursor cells (NPCs) and terminal differentiated neurons was performed according to previously described protocols with slight modifications[16,30,32]. All cell lines for each individual experiments were differentiated in parallel to further reduce experimental variability. Briefly, human pluripotent stem cell colonies were harvested using 1.5 mg $ml^{-1}$ collagenase type IV (Invitrogen), separated from the MEF feeder cells by gravity, gently triturated and cultured for 8 days in non-adherent suspension culture dishes (Corning) in EB medium (DMEM (Invitrogen) supplemented with 20% KnockOut Serum Replacement (Invitrogen), 0.5 mM glutamine (Invitrogen), 1% nonessential amino acids (Invitrogen), 0.1 mM β-mercaptoethanol (Sigma)) supplemented with 50 ng $ml^{-1}$ human recombinant Noggin (Peprotech) and 1,000 nM dorsomorphin (Stemgent). Subsequently human EBs were plated onto poly-L-ornithine (15 μg $ml^{-1}$, Sigma), laminin (1 μg $ml^{-1}$ Sigma), fibronectin (2 μg $ml^{-1}$ Sigma) coated tissue culture dishes in N2 medium[33] supplemented with 50 ng $ml^{-1}$ human recombinant Noggin (Peprotech), 1,000 nM dorsomorphin (Stemgent) and FGF2 (20 ng $ml^{-1}$, R&D systems). After 8 days, neural rosette-bearing EBs were cut out by microdissection, dissociated using 0.05% trypsin/EDTA solution (Invitrogen) and subsequently expanded on poly-L-ornithine, laminin and fibronectin coated cell culture dishes a density of $5 \times 10^5$ cells per $cm^2$ in N2 medium supplemented with FGF2 (20 ng $ml^{-1}$, R&D systems). Proliferating NPCs were passaged 2 to 4 times before induction of terminal differentiation into neurons by growth factor withdrawal in N2 medium supplemented with ascorbic acid (Sigma). Differentiated neurons were used for analysis between day 25 and 31 day after differentiation. After terminal differentiation, the cultures consist primarily of excitatory glutamatergic neurons and astrocytes with few other detectable cell types such as dopaminergic neurons (Extended Data Fig. 2). In addition neural precursor cells were also included in in gene expression analysis due to the robust expression of *SNCA* at this developmental stage.

**CRISPR/Cas9 gRNA and donor vector design.** To generate gRNA expression vectors, which express a fluorescent marker protein for FACS sorting in addition to Cas9 and the gRNA, we modified the pX330 gRNA expression vector[34] by insertion of either a CMV-EGFP-pA (pX330-EGFP), CMV-YFP-pA (pX330-YFP), CMV-mCherry-pA (pX330-mCherry) or CMV-BFP-pA (pX330-BFP) into the Not1 and Sbf1 restriction sites. Annealed oligonucleotides for each targeting site (Supplementary Table 3a) were ligated in to the BbsI restriction site as described previously[35]. Donor plasmids were generated by inserting a genomic PCR-amplified fragment (Supplementary Table 3b, 1,834 bp for the intron-4 enhancer and 2,333 bp for *SNCA-Rep1*) into the pCR2.1-TOPO-TA cloning vector (Life technologies) according to the provider's instructions. Donor plasmids for intron-4 enhancer targeting to insert the four genotypes (Supplementary Table 3b) were generated by replacing the wild-type sequence between the Mfe1 and Sty1 restriction sites with a synthetized gene fragment (gBlock, Integrated DNA technologies, Iowa). Donor plasmids for *SNCA-Rep1* targeting to insert the four *SNCA-Rep1* genotypes (Supplementary Table 3b) were generated by replacing the wild-type sequence between the Mfe1 and Stu1 restriction sites with corresponding *SNCA-Rep1* fragments selected from a sub-cloned haplotypes derived from a collection of human cell lines.

**CRISPR/Cas9-mediated genome editing of human ES cells.** CRISPR/Cas9-mediated genome editing of human ES cells was performed as described previously[30,36]. Human ES cells or the respective targeted sub-clones were cultured in Rho-associated protein kinase (ROCK)-inhibitor (10 μM, Stemgent; Y-27632) 24 h before electroporation. Cells were harvested using 0.05% trypsin/EDTA solution (Invitrogen) and resuspended in phosphate buffered saline (PBS). For genomic deletions, $1 \times 10^7$ cells were electroporated with 22.5 μg of each gRNA expression vector (Supplementary Table 3a). For insertion of *SNCA-Rep1* or intron-4 enhancer haplotypes, $1 \times 10^7$ cells were electroporated with 15 μg gRNA expression vector and 30 μg of the respective donor vector (Supplementary Table 3a, b). Cells were maintained on MEF feeder layers for 72 h in the presence of ROCK inhibitor followed by FACS sorting (FACS-Aria; BD-Biosciences) of a single-cell suspension for cells expressing the respective fluorescent marker proteins (Supplementary Table 3a) and subsequently plated at a low density in human ES cell medium supplemented with ROCK inhibitor for the first 24 h. Individual colonies were picked and expanded 10 to 14 days after electroporation. Correctly targeted clones were subsequently identified by genomic sequencing with primers outside of the targeting region (Supplementary Table 3c), by Southern blot analysis and for the *SNCA-Rep1* targeting by fragment length analysis. All targeted and maintained cell lines used for subsequent experiments are summarized in Supplementary Table 3g.

**Southern blotting.** Genomic DNA was separated on a 0.7% agarose gel after restriction digests with the appropriate enzymes, transferred to a nylon membrane (Amersham) and hybridized with $^{32}P$ random primer (Agilent)-labelled probes.

**Fragment length analysis.** Fragment length analysis for *SNCA-Rep1* was performed as described previously[26] using the Type-it Microsatellite PCR Kit (Qiagen) according to providers instruction using a *SNCA-Rep1*-forward and 5′-FAM (Fluorescein) labelled *SNCA-Rep-1* reverse Primer (Supplementary Table 3c). Fragment length along with an appropriate standard was determined using a 3730xl DNA analyser (Applied Biosystems) and analysed with Peak Scanner software 2 (Applied Biosystems).

**Genomic sequencing-based phase-reconstruction.** Genomic DNA from wild-type WIBR3 and targeted clones was amplified using PlatinumTaq DNA polymerase (Life Technologies) and primer pairs indicated in Supplementary Table 3d. PCR products were sub-cloned using TOPO XL-PCR cloning kit (Life Technologies) according to the providers' instructions and between 6 and 10 individual clones were submitted for sequencing. The phase between intron-4 enhancer SNPs (rs356168 and rs3756045) and the reporter SNP (rs356165) was manually determined based on the genotype of linked heterozygous SNPs in the respective fragments (Supplementary Table 3d).

**Allele-specific quantitative reverse transcription polymerase chain reaction (qRT–PCR).** RNA was isolated using the RNeasy Mini Kit (Qiagen) including on-column DNase digest to remove genomic DNA. Reverse transcription was performed on 0.5–1 μg of total RNA using oligo dT priming and SuperScript III First-Strand Synthesis SuperMix (Life technologies) at 50 °C according to the provider's instructions. The SNP genotyping TaqMan assays for allele-specific *SNCA* probes and TaqMan gene expression analysis assay targeting the 3′ UTR of *SNCA* and GAPDH were custom designed or provided by the manufacturer (Supplementary Table 3c; Applied Biosystems). All samples were performed in technical triplicates in a 384-well plate format on a 7900HT Fast Real-Time PCR system (Applied Biosystems) using Taqman Universal Master MIX II with UNG (Applied Biosystems) according to the manufacturers' instructions. Sequential dilution samples from neurons derived from human iPS cells that are homozygous for either the A allele (IPSC-PDA derived from AG20443)[16] or G allele (IPSC-PDB derived from AG20442)[16] at the reporter SNP (rs357165) were included to determine primer efficiency for each experiment separately. Relative quantification of allele-specific expression of *SNCA* was calculated using the Pfaffl method[37] that incorporates primer efficiencies, using total *SNCA* expression (3′ UTR-*SNCA* TaqMan assay) as the reference gene and dependent on experimental design, the average of either wild-type or enhancer deleted cells (WIBR3 ΔE4/ΔE4) as calibrator. Allele-specific *SNCA* expression proportions were calculated relative to calibrator samples which were set to 0.5. All statistical analyses were calculated using GraphPad Prism 6 for Mac. No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment. Significant differences between genotypes were calculated using one-way or two-way ANOVA followed by Tukey's multiple comparison test based on allele-specific expression from 3 biological replicates (representing the mean of 3 technical replicates) for each of the individual subclones as indicated in the figure legends. For statistical analysis all biological replicates of individual clones were combined according to genotypes. Equal variances between genotypes were tested using a Brown–Forsythe test. In addition the observed effects were confirmed using the non-parametric Kruskal–Wallis test followed by Dunn's multiple comparisons test. The allele-specific expression data and the detailed statistical analysis corresponding to data displayed in Fig. 2c, d and Extended Data Fig. 5d–g are provided as Source Data for Fig. 2 and Source Data for Extended Data Fig. 5.

**Reverse transcription of total RNA and real-time PCR.** RNA was isolated using the RNeasy Mini Kit (Qiagen) including on-column DNase digest to remove genomic DNA. Reverse transcription was performed on 0.5–1 μg of total RNA using oligo dT priming and SuperScript III First-Strand Synthesis SuperMix

(Life technologies) at 50 °C according to provider's instructions. All PCR reactions were performed in a 384-well plate format on a 7A900HT Fast Real-Time PCR system (Applied Biosystems) using either Taqman Universal Master MIX II with UNG (Applied Biosystems) or Platinum SYBR green qPCR SuperMIX-UDG with ROX (Invitrogen) according to the manufacturers' instructions using primer as summarized in Supplementary Table 3c. Relative quantification of gene expression was calculated using the $2^{-\Delta\Delta C_t}$ method using *GAPDH* (Taqman) or 60S acidic ribosomal protein P0 (*RPLP0* for SYBR green) as reference and untreated control samples (as specified in the figure legends) as calibrator. All statistical analyses were calculated using GraphPad Prism 6 for Mac. Detailed statistical information is included in the corresponding figure legends. No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

**Chromatin immunoprecipitation (ChIP-qRT–PCR and ChIP-seq).** ChIP-qRT–PCR and ChIP-seq was performed as described previously[38]. $10^7$ cells per ChIP assay were cross-linked for 10 min at room temperature by the addition of one-tenth of the volume of 11% formaldehyde solution (11% formaldehyde, 50 mM HEPES pH 7.3, 100 mM NaCl, 1 mM EDTA pH 8.0, 0.5 mM EGTA pH 8.0) to the growth media followed by quenching with 100 mM glycine. Cells were washed twice with PBS, then the supernatant was aspirated and the cell pellet was flash frozen in liquid nitrogen. 20 μl of magnetic Dynabeads (Sigma) were blocked with 0.5% BSA (w/v) in PBS. Magnetic beads were bound with 2 ug of antibody indicated in Supplementary Table 3e. Cross-linked cells were lysed with lysis buffer (50 mM HEPES pH 7.3, 140 mM NaCl, 1 mM EDTA, 10% glycerol, 0.5% NP-40, and 0.25% Triton X-100) and resuspended and sonicated in sonication buffer (50 mM Tris-HCl (pH 7.5), 140 mM NaCl, 1 mM EDTA, 1% Triton X-100, 0.1% Na-deoxycholate, 0.1% SDS). Cells were sonicated at 4 °C with a Bioruptor (Diagenode) at high power for 25 cycles for 30 s with 30 s between cycles. Sonicated lysates were cleared and incubated overnight at 4 °C with magnetic beads bound with antibody (Supplementary Table 3e) to enrich for DNA fragments bound by the indicated TF. Beads were washed two times with sonication buffer, one time with sonication buffer with 500 mM NaCl, one time with LiCl wash buffer (20 mM Tris pH 8.0, 1 mM EDTA, 250 mM LiCl, 0.5% NP-40, 0.5% Na-deoxycholate) and one time with TE with 50 mM NaCl. DNA was eluted in elution buffer (50 mM Tris-HCL pH 8.0, 10 mM EDTA, 1% SDS). Cross-links were reversed overnight. RNA and protein were digested using RNase A and Proteinase K, respectively and DNA was purified with phenol chloroform extraction and ethanol precipitation. Target-specific binding was analysed by quantitative RT–PCR on a 7900HT Fast Real-Time PCR system (Applied Biosystems) using with Platinum SYBR green pPCR SuperMIX-UDG with ROX (Invitrogen) using primers targeting either the intron-4 enhancer or an adjacent negative control in the *SNCA* locus or a unrelated negative control region on chromosome 8. Target-specific binding of the intron-4 enhancer or control region for each antibody was calculated as fold-enrichment over IgG-Isotype control ChIP. Similarly processed H3K27ac samples post-purification were used to prepare Illumina multiplex sequencing libraries. Libraries for Illumina sequencing were prepared following the Illumina TruSeq DNA Sample Preparation v2 kit protocol with the following exceptions. After end-repair and A-tailing, immunoprecipitated DNA (~10–50 ng) or whole cell extract DNA (50 ng) was ligated to a 1:50 dilution of Illumina Adaptor Oligo Mix assigning one of 24 unique indexes in the kit to each sample. Following ligation, libraries were amplified by 18 cycles of PCR using the HiFi NGS Library Amplification kit from KAPA Biosystems. Amplified libraries were then size-selected using a 2% gel cassette in the Pippin Prep system from Sage Science set to capture fragments between 300 and 700 bp. Libraries were quantified by qRT–PCR using the KAPA Biosystems Illumina Library Quantification kit according to kit protocols. Libraries with distinct TruSeq indexes were multiplexed by mixing at equimolar ratios and running together in a lane on the Illumina HiSeq 2500 for 40 bases in single read mode. Reads were de-multiplexed by their adaptor ID and aligned to the hg19 version of the human reference genome using Bowtie 1.1.1 with parameters –wrapper basic-0 –p 4 –k 2 –m 2 –sam and –l 40 (ref. 39).

**Parkinson's disease associated SNP prioritization based on overlap with epigenetic enhancer marks.** We prioritized Parkinson's disease associated high-confidence enhancer elements as being broadly active throughout the brain based on the observation that Parkinson's disease associated pathology is not exclusively localized to the substantia nigra, but rather affects a wide variety of neuronal cell types[40–42]. DHSs data sets were included because the comparison between fetal and adult DHSs harbouring disease-associated GWAS variants suggests that the vast majority of these DHSs are already established in fetal tissues[3]. Mapped read files of H3K27ac ChIP-seq, H3K4me1 ChIP-seq, H3K4me3 ChIP-seq, and DHSs created as part of the Human Epigenome Roadmap Project were downloaded from Gene Expression Omnibus (GEO), and the samples used are summarized in Supplementary Table 3f. These data sets were used to prioritize Parkinson's disease associated SNPs by their presence in putative regulatory regions in human brain samples (Supplementary Table 1). Regions of enrichment in H3K27ac signal, H3K4me1 signal, H3K4me3, or DHSs in human samples were calculated using MACS 1.4.2 (ref. 43) using parameters -p 1e-9,–keep-dup = auto, and -g hs with no control library. For each SNP in the *SNCA* locus from hg19 chromosome 4 position 90453241 (rs67262058) to position 91220981 (rs17016715) (463 SNPs, $P < 5 \times 10^{-8}$ provided by PDGene database[12,44]), the number of data sets with a region enriched in H3K4me1, H3K27ac, or DNase hypersensitivity contacting or containing the SNP were summed. SNPs were ranked by this sum. SNPs in the promoter region of *SNCA* were excluded based the overlap of H3K4me3 peaks, a histone modification that is enriched at the transcription start site (TSS) of actively expressed genes identified in each ChIP-seq sample.

**Display of ChIP-seq and DHSs signal.** WIG files for display of read pileup for ChIP-seq and DHSs performed by the Epigenome Roadmap Consortium were downloaded from GEO, and these samples are compiled in Supplementary Table 3f. For new data sets generated for this manuscript we created WIG files representing read counts in 50 bp bins using macs 1.4 (ref. 43) with parameters –w –S –space = 50 –nomodel –shiftsize –p 1e-9 –keep-dup = 1. Samples were normalized to reads in each bin per million mapped reads to enable cross-sample comparisons. This was achieved by dividing the value in each bin by the millions of mapped reads per sample and conversion to TDF files for display in IGV using igvtools.

**Relative enhancer ChIP-seq signal calculation in CRISPR/Cas9 targeted cells.** For Extended Data Fig. 7e, the signal of H3K27ac ChIP-seq at the enhancer containing the variable allele was calculated in five cell lines with and without deletion or additional alteration of the underlying sequence. The SNP-containing enhancer was defined as chr4: 90673430–90675431, and the 3′ UTR enhancer used for normalization was defined as chr4: 90629326–90631326. Aligned reads in these windows were counted using intersectBed[45] Because visual inspection indicated that the 3′ UTR enhancer was consistent in signal, and because no genetic perturbation to this region was performed, we used the read count in this region to normalize across samples for comparison. The ratio of number of reads in the variable SNP-containing enhancer divided by reads in the 3′ UTR enhancer is reported.

**Conditional GWAS meta-analysis.** Genomewide SNP data was obtained for 5 cohorts (GenePD/PROGENI, phs000126.v1.p1; NINDS, phs000089.v3.p2; NGRC, phs000196.v2.p1; APDGC, phs000394.v1.p1; WTCCC, EGAS00000000034). Each cohort was imputed to the 1000G phase 3 reference genome using Minimac[46,47]. Baseline and conditional association of the GWAS hits reported by Nalls *et al.* to Parkinson's disease risk was performed using SAS for each of the 5 cohorts as describe previously[12]. Association results for each cohort were combined using fixed effects models with standard error weighting implemented in METAL[48].

**eQTL study in Parkinson's disease and control postmortem frontal cortex.** The *SNCA* gene expression data set and analysis methods used here have been described previously[49]. In brief, *SNCA* expression levels were assayed from 165 samples using quantitative real-time polymerase chain reaction. The relative standard curve method was used to transform the $C_t$ values into quantity units. The base 10 logarithm of the *SNCA* expression values was used for all analyses, to ensure the normal distribution of data required by the statistical tests performed. SNP genotyping was performed as part of the US PD-GWAS Consortium meta-analysis replication sample[50]. As described in the consortium study, the samples were genotyped at the Center for Inherited Disease Research (CIDR) using a custom Illumina genotyping array of 768 SNPs. Because the tested SNP data set did not include the top reported Parkinson's disease associated SNP rs356182 in the *SNCA* locus, we included rs356229 as proxy for this SNP ($R^2 = 0.62$) in our analysis. After quality control, genotyping and expression data was available for 86 cases and 41 controls for eQTL analysis. Expression models were analysed including adjustment for disease status, sex, pH, age at death, as well as for the interaction between PMI and disease status and significance was assessed using a one-sided test based on the *a priori* hypothesis of an association of the G allele at rs356168 with increased *SNCA* expression.

**TF binding site prediction.** We extracted the nucleotide sequence for the region chr4: 90450000–91221000, that includes the *SNCA* locus, using the hg19 genome release, and made an alternative 'minor allele' version of the sequence that contains all Parkinson's disease associated SNP-derived minor alleles from position 90453241 (rs67262058) to position 91220981 (rs17016715). Any minor allele SNPs less than 50 nt apart were included in separate sequence files. This way we could evaluate the effect of the minor allele within the reference wild-type sequence context. We predicted all TF binding sites for the reference sequence and the sequences containing minor alleles using match, the matrix library "matrix.dat" from the TRANSFAC Release 2014.1, and the matrix profile minSUM_good.prf. We summarized the TFs predicted to bind the wild-type and minor allele sequences with the BEDTools suite and a custom perl script. To select candidate TFs, the list of differential binding matrices was filtered for TFs that (i) show high expression

levels in the top 25th percentile of all transcripts based on RNA-sequencing data derived from *in vitro* differentiated neurons (data not shown), (ii) show robust expression in disease-relevant brain areas based on data from the Allen Brain Atlas[51] and (iii) display gene ontology (GO) terms related to brain function[52,53].

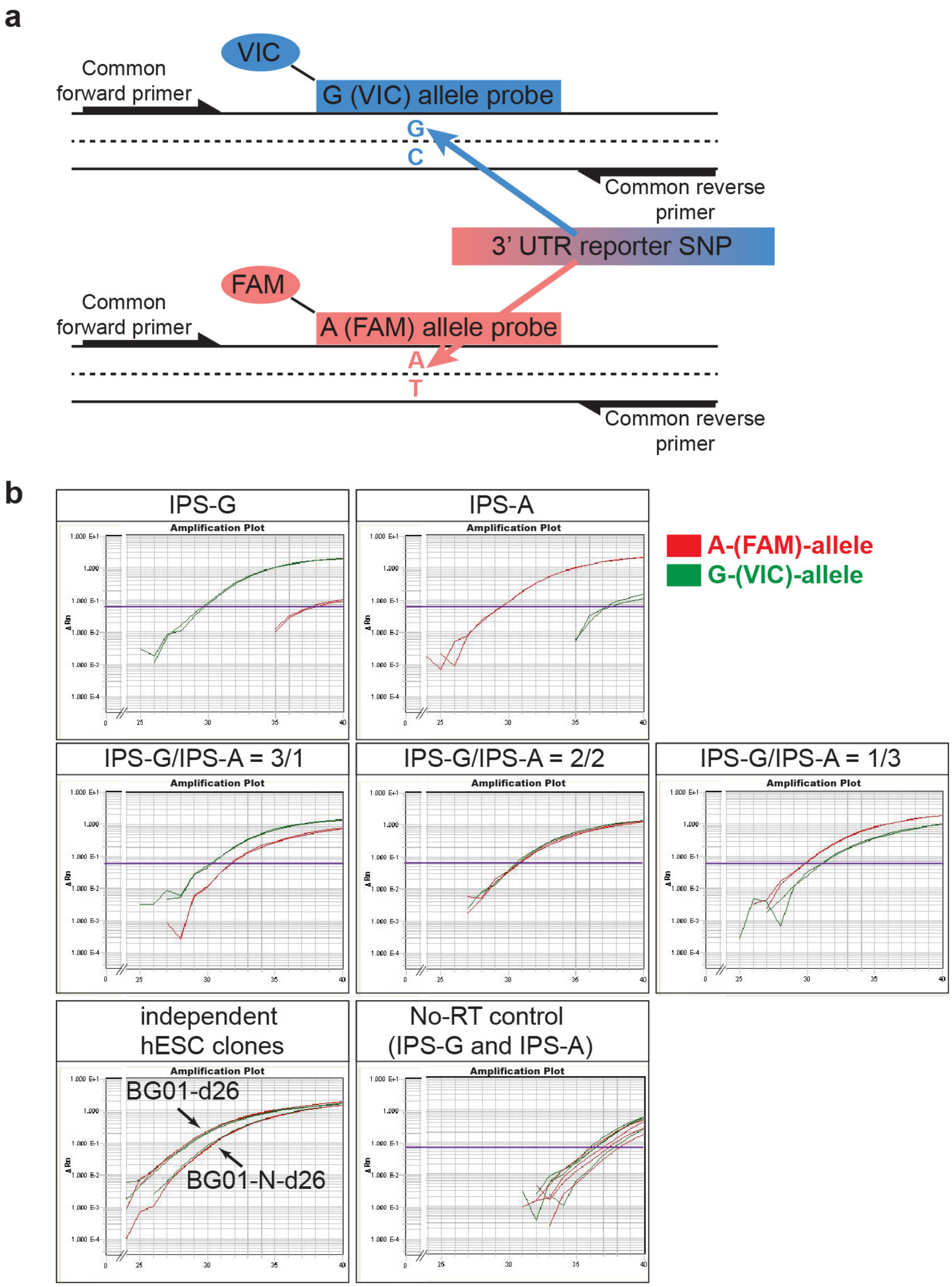
**Electrophoretic mobility shift assay (EMSA).** HEK-293 cells transiently transfect using X-tremeGene 9 (Roche) with plasmids expressing either EMX2–Myc-DDK or NKX6-1-Myc-DDK (pCMV6-Entry, OriGene) or wild-type cells were used to prepare nuclear extracts according to standard protocols. Gel-shift assay was performed with the LightShift Chemiluminescent EMSA Kit (Thermo Scientific) according to manufacturer's instructions. For the competition experiments, 5′-biotin-labelled oligonucleotides carrying either the A allele or G allele at rs356168 spanning 60 bp around the SNP and corresponding unlabelled competitor oligonucleotides (all Integrated DNA Technologies) were included in the binding reaction to determine the allele-specific binding of EMX2 or NKX6-1 respectively. Binding reactions were separated using precast 6% DNA retardation gels or 4–20% TBE gels (Life Technologies) in 0.5X TBE, electrophoretically transferred to Biodyne B nylon membranes (Thermo Scientific) and detected by chemiluminescence.

**Doxycycline-inducible expression of EGFP, EMX2 and NKX6-1 in human ES-cell-derived neurons.** Myc-DDK-tagged human cDNAs for EMX2 and NKX6-1 (derived from pCMV6-Entry Clones, OriGene) were subcloned into FUW-Tet-O-EGFP vectors to generate the FUW-Tet-O-EMX2 and FUW-Tet-O-NKX6-1 doxycycline-responsive lentiviral vectors. VSV-G coated lentiviruses were generated in 293T cells as described previously[54]. Briefly, 293T cells were transfected with a mixture of viral plasmid and packaging constructs expressing the viral packaging functions and the VSV-G protein. Culture medium was changed 12 h post-transfection and virus-containing supernatant was collected 60–72 h post transfection. Viral supernatant was filtered through a 0.45 μm filter. Virus-containing supernatants were concentrated by ultracentrifugation. All experiments were performed in 2 individual clones derived from WIBR3 cells, in which the constitutive active reverse tetracycline transactivator (AAVSI-neo-M2rtTA) was targeted into the AAVSI 'safe harbour' locus as described previously[36,55]. *In vitro* differentiated neural precursors derived from these cells lines were transduced with various amounts of concentrated virus for 24 h. Only cultures that expressed the transgenes in more than 75% of the cells (as determined by GFP fluorescence and immunostaining for expression of the DDK-tag) were used for further experiments. Transduced cells were terminally differentiated for 21 days. Transgene expression was induced in 24-well plates by supplementing the medium with doxycycline at a final concentration of 2 μg $ml^{-1}$. Doxycycline-induced transgene expressing samples and untreated isogenic controls were lysed 72 h after doxycycline and subsequently analysed by qRT–PCR.

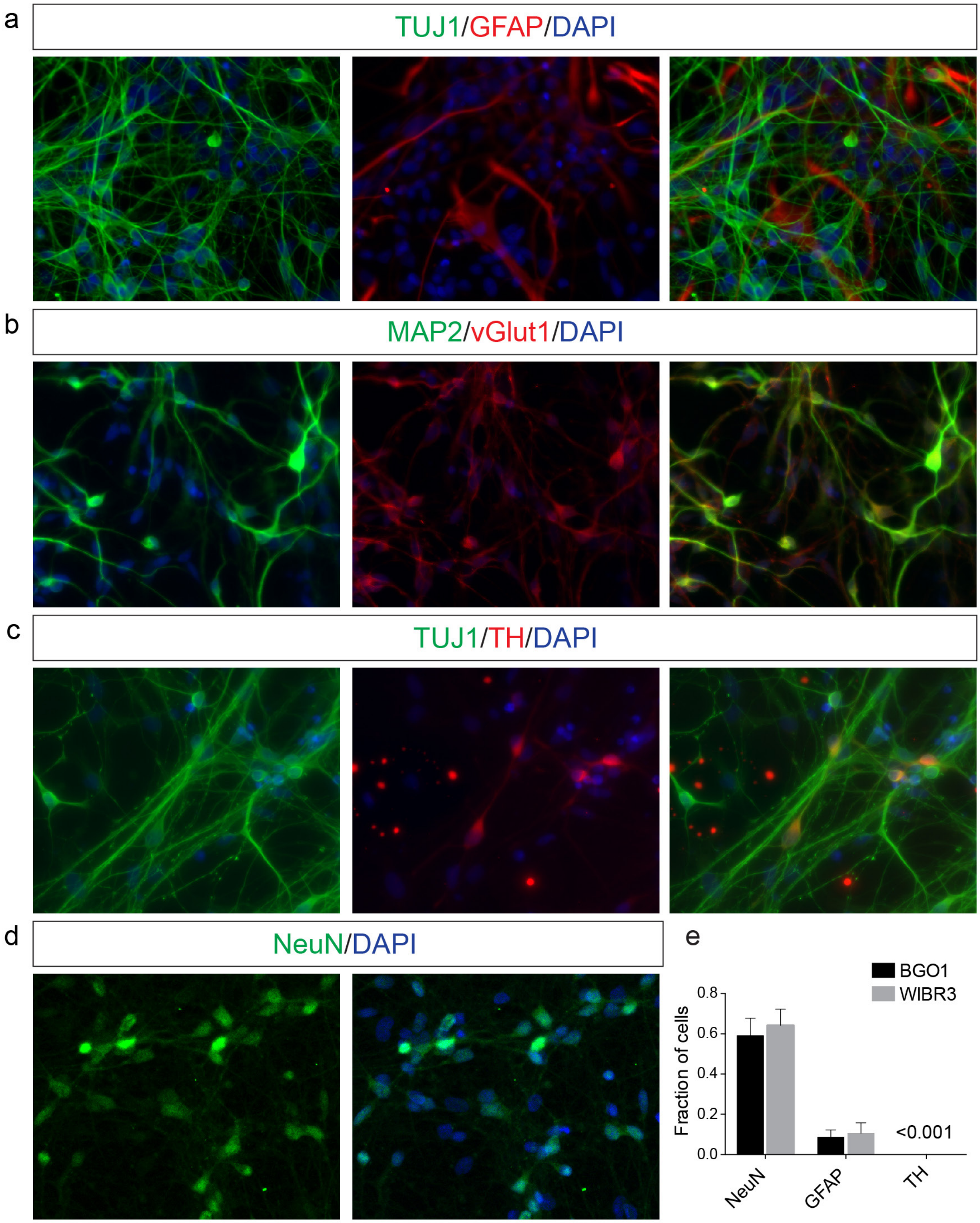
**Single molecule mRNA flouresence *in situ* hybridization (FISH).** We performed RNA FISH as outlined previously[56–58]. The neuronal cultures (differentiation day 21) were washed with HBSS, detached and dissociated into single cells with HBSS and subsequently fixed with paraformaldehyde at a final concentration of 4%. The cells were incubated for 10 min while rotating to avoid clumping of cells. After 10 min the cells were spun down for 6 min at 1,000 r.p.m. To permeablize the cells, the cells were placed in 70% ethanol overnight. The next day the cells were attached to chambered cover slides (Nunc Lab-Tek) coated with poly-L-lysine. The 20 nt probes for EMX2 and NKX6-1 were manually designed and ordered through Biosearch Technologies, coupled with either Cy5-flurophore or Alexa594-fluorophore (Invitrogen), respectively, and hybridized with standard FISH hybridization buffer containing 25% formamide. For hybridization conditions 75 ng probes per μl of hybridization buffer were used. The probes were hybridized for 16 h at 30 °C followed by two wash steps with wash buffer containing 25% formamide and 2× SSC. The cells were counterstained with Hoechst 33342. During imaging the cells were kept in a solution containing PBS, glucose, catalase and Trolox to avoid bleaching of fluorophores. All images were taken with a Nikon Ti-E inverted fluorescence microscope equipped with a 100× oil-immersion objective and a Photometrics Pixis 1024 CCD camera using MetaMorph software (Molecular Devices, Downington, PA). A total of 100 cells were counted for quantification. Cells that seemed fragmented or had an excessive amount of background were excluded. All cells with more than two transcripts for either NXK6-1 or EMX2 were counted as positive for the respective transcript.

**Immunocytochemistry.** Cells were fixed with 4% (w/v) paraformaldehyde in PBS for 20 min at room temperature, and rinsed with PBS. Following membrane permeabilization with PBS containing 0.2% Triton, cells were blocked with 5% normal donkey serum and stained with indicated primary antibodies (Supplementary Table 3e) overnight at 4 °C. Immunostainings were visualized by appropriate secondary antibodies conjugated with Alexa 488, 568, 594, 633 (Life Technologies), followed by counter-staining with DAPI.

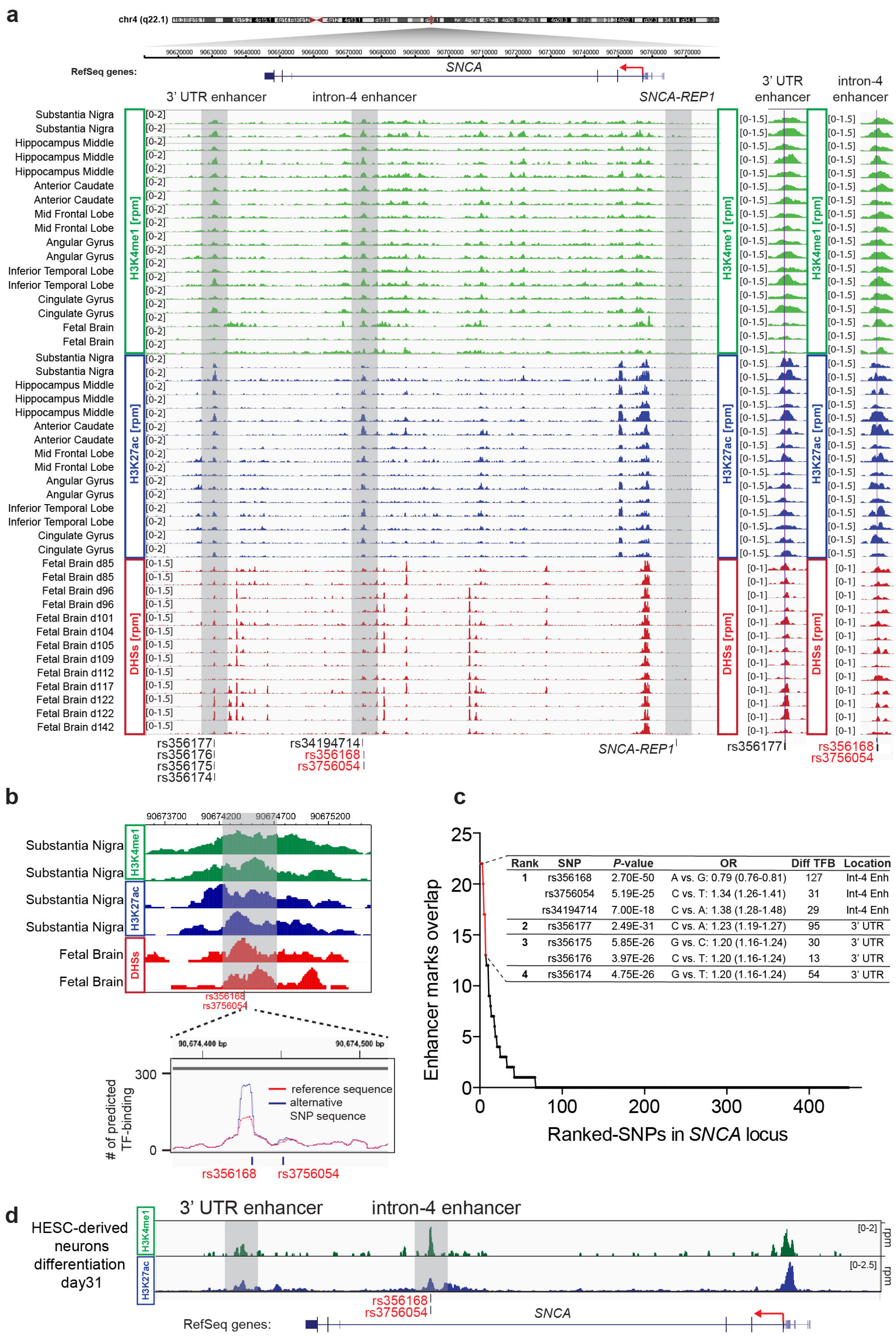

31. Lengner, C. J. *et al.* Derivation of pre-X inactivation human embryonic stem cells under physiological oxygen concentrations. *Cell* **141,** 872–883 (2010).
32. Kim, J. *et al.* Direct reprogramming of mouse fibroblasts to neural progenitors. *Proc. Natl Acad. Sci. USA* **108,** 7838–7843 (2011).
33. Kim, J.-H., Panchision, D., Kittappa, R. & McKay, R. Generating CNS neurons from embryonic, fetal, and adult stem cells. *Methods Enzymol.* **365,** 303–327 (2003).
34. Cong, L. *et al.* Multiplex genome engineering using CRISPR/Cas systems. *Science* **339,** 819–823 (2013).
35. Wang, H. *et al.* One-step generation of mice carrying mutations in multiple genes by CRISPR/Cas-mediated genome engineering. *Cell* **153,** 910–918 (2013).
36. Hockemeyer, D. *et al.* Efficient targeting of expressed and silent genes in human ESCs and iPSCs using zinc-finger nucleases. *Nature Biotechnol.* **27,** 851–857 (2009).
37. Pfaffl, M. W. A new mathematical model for relative quantification in real-time RT–PCR. *Nucleic Acids Res.* **29,** e45 (2001).
38. Lee, T. I., Johnstone, S. E. & Young, R. A. Chromatin immunoprecipitation and microarray-based analysis of protein location. *Nature Protocols* **1,** 729–748 (2006).
39. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10,** R25 (2009).
40. Sulzer, D. & Surmeier, D. J. Neuronal vulnerability, pathogenesis, and Parkinson's disease. *Mov. Disord.* **28,** 41–50 (2013).
41. Ferrer, I., Martinez, A., Blanco, R., Dalfó, E. & Carmona, M. Neuropathology of sporadic Parkinson disease before the appearance of parkinsonism: preclinical Parkinson disease. *J. Neural Transm.* **118,** 821–839 (2011).
42. Irwin, D. J. *et al.* Neuropathologic substrates of Parkinson disease dementia. *Ann. Neurol.* **72,** 587–598 (2012).
43. Zhang, Y. *et al.* Model-based analysis of ChIP-seq (MACS). *Genome Biol.* **9,** R137 (2008).
44. Lill, C. M. *et al.* Comprehensive research synopsis and systematic meta-analyses in Parkinson's disease genetics: the PDGene database. *PLoS Genet.* **8,** e1002548 (2012).
45. Quinlan, A. R. BEDTools: the Swiss-Army tool for genome feature analysis. *Curr Protoc. Bioinformat.* **47,** 11.12.1–11.12.34 (2014).
46. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G. R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nature Genet.* **44,** 955–959 (2012).
47. Fuchsberger, C., Abecasis, G. R. & Hinds, D. A. minimac2: faster genotype imputation. *Bioinformatics* **31,** 782–784 (2015).
48. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26,** 2190–2191 (2010).
49. Dumitriu, A. *et al.* Cyclin-G-associated kinase modifies α-synuclein expression levels and toxicity in Parkinson's disease: results from the GenePD study. *Hum. Mol. Genet.* **20,** 1478–1487 (2011).
50. Pankratz, N. *et al.* Meta-analysis of Parkinson's disease: identification of a novel locus, *RIT2*. *Ann. Neurol.* **71,** 370–384 (2012).
51. Hawrylycz, M. J. *et al.* An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature* **489,** 391–399 (2012).
52. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* **37,** 1–13 (2009).
53. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols* **4,** 44–57 (2009).
54. Brambrink, T. *et al.* Sequential expression of pluripotency markers during direct reprogramming of mouse somatic cells. *Cell Stem Cell* **2,** 151–159 (2008).
55. DeKelver, R. C. *et al.* Functional genomics, proteomics, and regulatory DNA analysis in isogenic settings using zinc finger nuclease-driven transgenesis into a safe harbor locus in the human genome. *Genome Res.* **20,** 1133–1142 (2010).
56. Raj, A., van den Bogaard, P., Rifkin, S. A., van Oudenaarden, A. & Tyagi, S. Imaging individual mRNA molecules using multiple singly labeled probes. *Nature Methods* **5,** 877–879 (2008).
57. Raj, A., Rifkin, S. A., Andersen, E. & van Oudenaarden, A. Variability in gene expression underlies incomplete penetrance. *Nature* **463,** 913–918 (2010).
58. Faddah, D. A. *et al.* Brief Report. *Cell Stem Cell* **13,** 23–29 (2013).



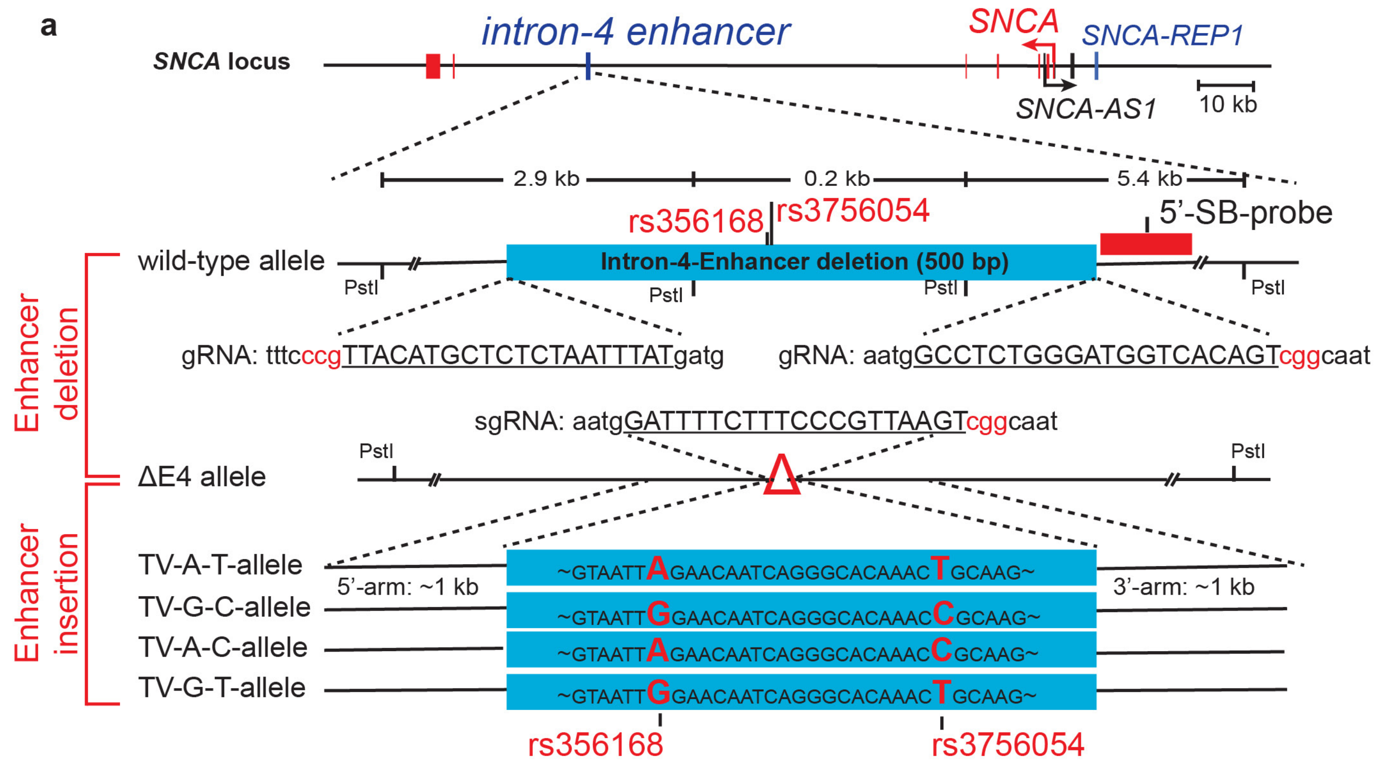

**Extended Data Figure 1 | Analysis of allele-specific expression of *SNCA* using qRT–PCR. a**, Schematic illustration of the quantitative allele-specific *SNCA* expression analysis using a common primer pair and allele-specific Taqman probes conjugated with different fluorophores to specifically detect a reporter-SNP (rs356165) in the 3′ UTR of *SNCA* in a multiplex reaction. As indicated, 6-carboxyfluorescein (FAM) and 4,7,2′-trichloro-7′-phenyl-6-carboxyfluorescein (VIC) were used to detect the A and G allele respectively. **b**, Representative multiplex qRT–PCR reactions (in duplicates) measuring allele-specific *SNCA* expression of allele-biased samples described in Fig. 1c, d. Allele-biased samples were generated by mixing human iPS-cell-derived neurons homozygous for either the A (IPS-A) or G allele (IPS-G) at rs356165 at indicated ratios. Also included is a plot showing cDNAs synthesized in the absence of SuperScript reverse transcriptase (no RT) to control for genomic DNA contaminations. Plots are displayed as reporter dye fluorescence signal ($\Delta$Rn) in log scale as a function of run cycle.

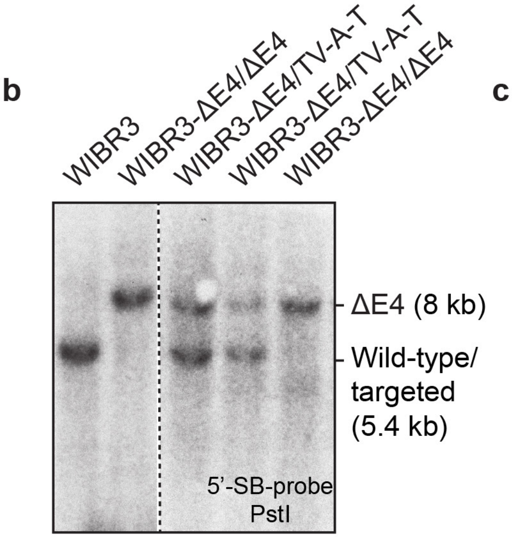



**Extended Data Figure 2 | Analysis of *in vitro* differentiated human ES-cell-derived mixed neuronal cultures. a–d**, Immunostainings of *in vitro* differentiated mixed neuronal cultures (differentiation day 28) for expression of neuron- and astrocyte-specific markers. Shown are representative images for staining of neuron-specific beta-III-tubulin (TUJ1) and astrocyte-specific glial fibrillary acidic protein (GFAP) (**a**), neuron-specific microtubule associated protein 2 (MAP2) and glutamatergic neuron-specific glutamate vesicular transporter 1 (vGLUT1) (**b**), TUJ1 and dopaminergic neuron-specific tyrosine-hydroxylase (TH) (**c**) and the pan-neuronal marker NeuN, which was used for quantification (**d**). **e**, Quantification of a representative *in vitro* differentiation experiment in human ES cell line WIBR3 and BGO1. The quantification of rare TH-positive cells was estimated. Source Data for this figure are available online.




**Extended Data Figure 3 | Identification of Parkinson's disease associated risk variants overlapping with distal enhancers in the *SNCA* locus. a**, Detailed H3K4me1 and H3K27ac ChIP-seq and DHSs-enrichment tracks for indicated brain regions in the *SNCA* locus. Shown are the locations of *SNCA-Rep1* and Parkinson's disease associated SNPs overlapping with two proximal enhancer elements (3′ UTR enhancer and intron-4 enhancer) highlighted by light grey boxes. On the right, enlarged view of 3′ UTR enhancer and intron-4 enhancer relative to top ranked Parkinson's disease associated SNPs. **b**, Enlarged view of the intron-4 enhancer region showing H3K4me1 and H3K27ac ChIP-seq enrichment tracks for substantia nigra and DHSs enrichment tracks for fetal brain relative to location of PD-associated SNPs. Shown below is the number of predicted TF binding sites for reference (in red) and alternative SNP (minor allele) sequence (in blue) at each genomic position. Grey box indicates location of deletion described in Fig. 2b and Extended Data Fig. 4a. **c**, Summary of all Parkinson's disease associated SNPs in the *SNCA* locus ranked by cumulative overlap with H3K4me1, H3K27ac and DHSs enhancer marks. Table summarizes the top 7 ranked SNPs with Parkinson's disease association *P* values, odd ratios (OR), number of predicted differential TF binding sites (Diff TFB) and location within enhancer elements as marked in **a**. **d**, Gene tracks showing H3K4me1 and H3K27ac enrichment in the *SNCA* locus for *in vitro* human ES-cell-derived neurons (differentiation day 31). hESC, human embryonic stem cell.

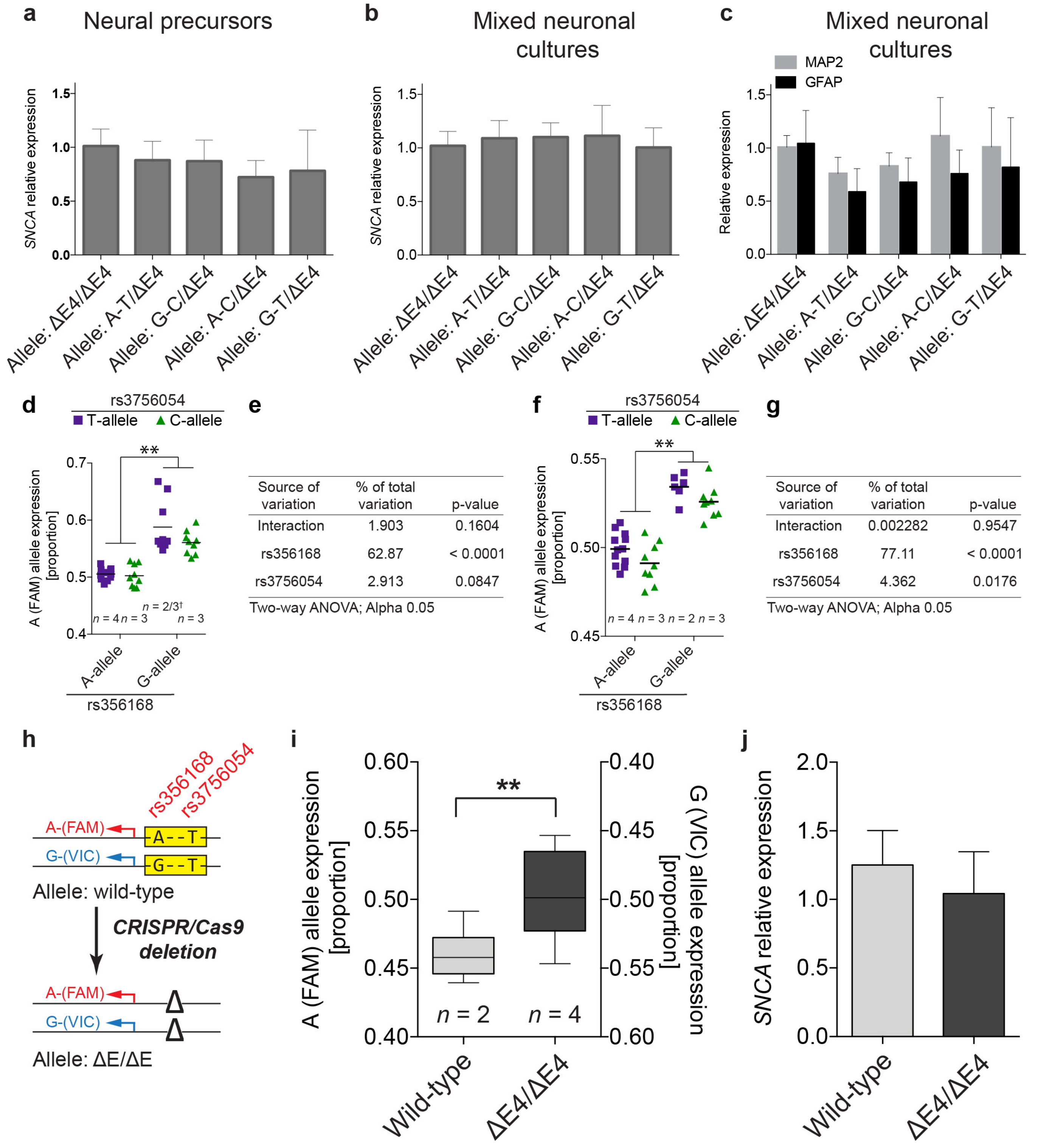



c

| Targeted modification | WIBR3 | In-phase with A (FAM) reporter SNP |
|---|---|---|
| Homozygous deletion (confirmed by sequencing) | 4 /48 (12%) (4/4 confirmed) | N/A |
| Heterozygous deletion (confirmed by sequencing) | 10/48 (20.1%) (0/10 confirmed) | N/A |
| Enhancer insertion (confirmed by sequencing and Southern blot) | TV-A-T: 8/96 (8.3%)<br>TV-G-C: 4/96 (4.2%)<br>TV-A-C: 5/96 (5.2%)<br>TV-G-T: 4/268 (1.5%) | 4/8<br>3/4<br>3/5<br>2/4 |

**Extended Data Figure 4 | CRISPR/Cas9-mediated genome editing strategy for targeted insertion of Parkinson's disease associated intron-4 enhancer elements in human ES cells. a**, Schematic illustration of the CRISPR/Cas9-mediated two-step genome editing strategy to delete and subsequently insert indicated intron-4 enhancer sequences containing the Parkinson's disease associated risk SNPs rs356168 and rs3756054. Shown are the genomic organization of the *SNCA* locus, an enlarged view of wild-type and deleted (ΔE4) alleles, gRNA targeting sequences (underlined, PAM sequence in red), restriction sites, Southern blot (SB) probe and design of targeting vectors (TV; risk SNPs are highlighted in red). **b**, Representative Southern blot analysis of indicated targeted WIBR3 human ES cells (ΔE4/TV-A-T) compared with wild-type cells or human ES cells carrying homozygous deletions (ΔE4/ΔE4). **c**, Table summarizing intron-4 enhancer deletions and insertions of indicated haplotypes in WIBR3 human ES cells. Correct targeting was determined by Southern blot analysis and genomic sequencing. Cell lines with targeted enhancer elements confirmed to be in *cis* with A (FAM) reporter SNP (determined by genomic sequencing-based phase-reconstruction) were maintained for subsequent analysis (compare Fig. 2b).




Panel e:

| Source of variation | % of total variation | p-value |
|---|---|---|
| Interaction | 1.903 | 0.1604 |
| rs356168 | 62.87 | < 0.0001 |
| rs3756054 | 2.913 | 0.0847 |

Two-way ANOVA; Alpha 0.05

Panel g:

| Source of variation | % of total variation | p-value |
|---|---|---|
| Interaction | 0.002282 | 0.9547 |
| rs356168 | 77.11 | < 0.0001 |
| rs3756054 | 4.362 | 0.0176 |

Two-way ANOVA; Alpha 0.05

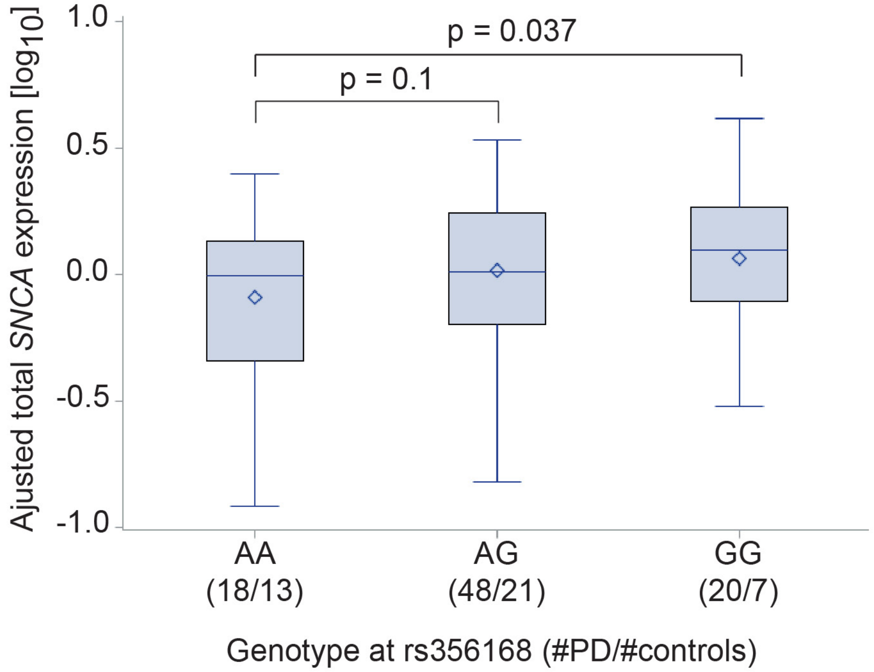
**Extended Data Figure 5** | See next page for caption.

**Extended Data Figure 5 | Effect of intron-4 enhancer modification on total and allele-specific expression of *SNCA*.** **a**, **b**, Analysis of total *SNCA* expression in *in vitro* derived neural precursors (**a**) (same samples analysed in Fig. 2c) and mixed neuronal cultures (**b**) (same samples analysed in Fig. 2d) from targeted cell lines with indicated SNP genotypes at rs356168 and rs3756054 (compare Fig. 2b) compared to human ES cells harbouring homozygous deletions of the intron-4 enhancer (ΔE4/ΔE4). qRT–PCR data are normalized to *GAPDH* and presented relative to the expression of ΔE4/ΔE4 cells. **c**, Expression analysis for the neuron-specific marker microtubule associated protein 2 (*MAP2*) and astrocyte specific glial fibrillary acidic protein (*GFAP*) by qRT–PCR in *in vitro* differentiated neurons described in **b**. Data are normalized to 60S acidic ribosomal protein P0 (*RPLP0*) and presented relative to the expression of ΔE4/ΔE4 neurons. Data are shown as mean ± s.d. of 3 biological replicates (each representing 3 technical replicates) for independent targeted clones as described in Fig. 2c, d ($n$ indicates number of independently targeted clones per genotype; ΔE4/ΔE4, $n = 4$; A–T/ΔE4, $n = 4$; G–C/ΔE4, $n = 3$; A–C/ΔE4, $n = 3$; G–T/ΔE4, $n = 2/3$). **d**, Alternative presentation of data displayed in Fig. 2c as dot blot grouped according to SNPs rs356168 and rs3756054 (excluding data for cells carrying homozygous deletions (ΔE4/ΔE4) of the intron 4 enhancer). **e**, Table summarizing the results of two-way ANOVA analysis corresponding to **d** for allele-specific *SNCA* expression in neural precursors. **f**, Alternative presentation of data displayed in Fig. 2d as dot blot grouped according to SNPs rs356168 and rs3756054 (excluding data for cells carrying homozygous deletions (ΔE4/ΔE4) of the intron-4 enhancer). Each dot represents mean of 3 technical replicates, black bars indicate mean for each genotype (**d**, **f**). **g**, Table summarizing the results of two-way ANOVA analysis corresponding to **f** for allele-specific *SNCA* expression in mixed neuronal cultures. Allele-specific expression for each clone was analysed in 3 independent biological replicate experiments and combined according to genotypes, $n$ indicates number of independently targeted clones per group, † indicates an additional sub-clone derived from one of the two targeted clones for this genotype. Two-way ANOVA analysis (alpha = 0.05) was calculated based on allele-specific expression of all biological replicates. *$P < 0.0001$. **h**, Schematic illustration of the experimental strategy for CRISPR/Cas9-mediated deletion of the intron-4 enhancer element. Genomic sequencing-based phase-reconstruction, to analyse the phase of the heterozygous enhancer SNP rs356168 in wild-type WIBR3 cells indicates that the functional G allele at rs356168 is in *cis* with the G allele of the reporter SNP rs356165. **i**, Relative allele-specific *SNCA* expression (boxplots showing median, 25th and 75th percentiles with whiskers indicating minimum and maximum) in neural precursors comparing wild-type cells to ΔE4/ΔE4 cells harbouring homozygous deletions (expression is calculated relative to ΔE4/ΔE4 neural precursors). Note that the expression of the A (FAM) allele (displayed on the left $y$ axis) is reciprocal to the expression of the G (VIC) allele (displayed on the right $y$ axis). Statistically significant differences between groups were calculated using unpaired two-tailed $t$-test. $n$ indicates number of independent sub-clones for wild-type cells or independently deleted clones for ΔE4/ΔE4 cells; allele-specific expression for each clone was analysed in 3 independent biological replicate experiments at passage 1 and 2 respectively, each measured as 3 technical replicates. **$P < 0.0001$. **j**, qRT–PCR expression analysis for total *SNCA* in same samples as described in **i**. Data are normalized to *GAPDH* and displayed relative to the expression of ΔE4/ΔE4 neural precursors. Data are displayed as mean ± s.d.; Source Data and detailed statistical analysis are provided online.

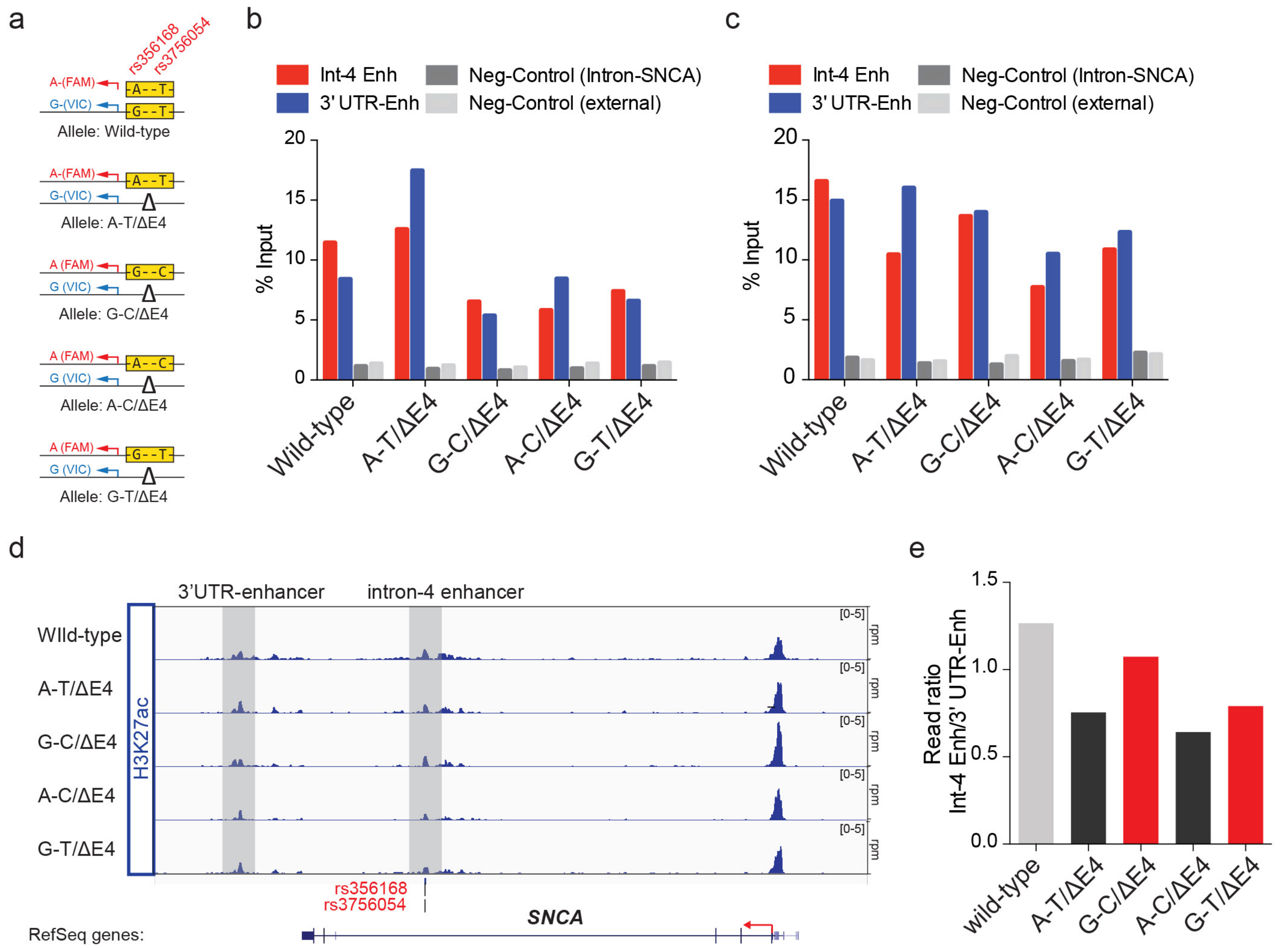
**a**

| | | | Unadjusted | | Conditioned on rs356168 | |
|---|---|---|---|---|---|---|
| **SNP** | **MAF** | **Allele** | **OR** | **p-value** | **OR** | **p-value** |
| rs356168 | 0.47 | G | 1.26 | 7.4e-18 | n/a | n/a |
| rs356182 | 0.37 | G | 1.32 | 1.1e-21 | 1.23 | 3.0e-6 |
| rs7681154 (conditional top hit) | 0.50 | C | 1.05 | 0.078 | 1.16 | 1.0e-7 |

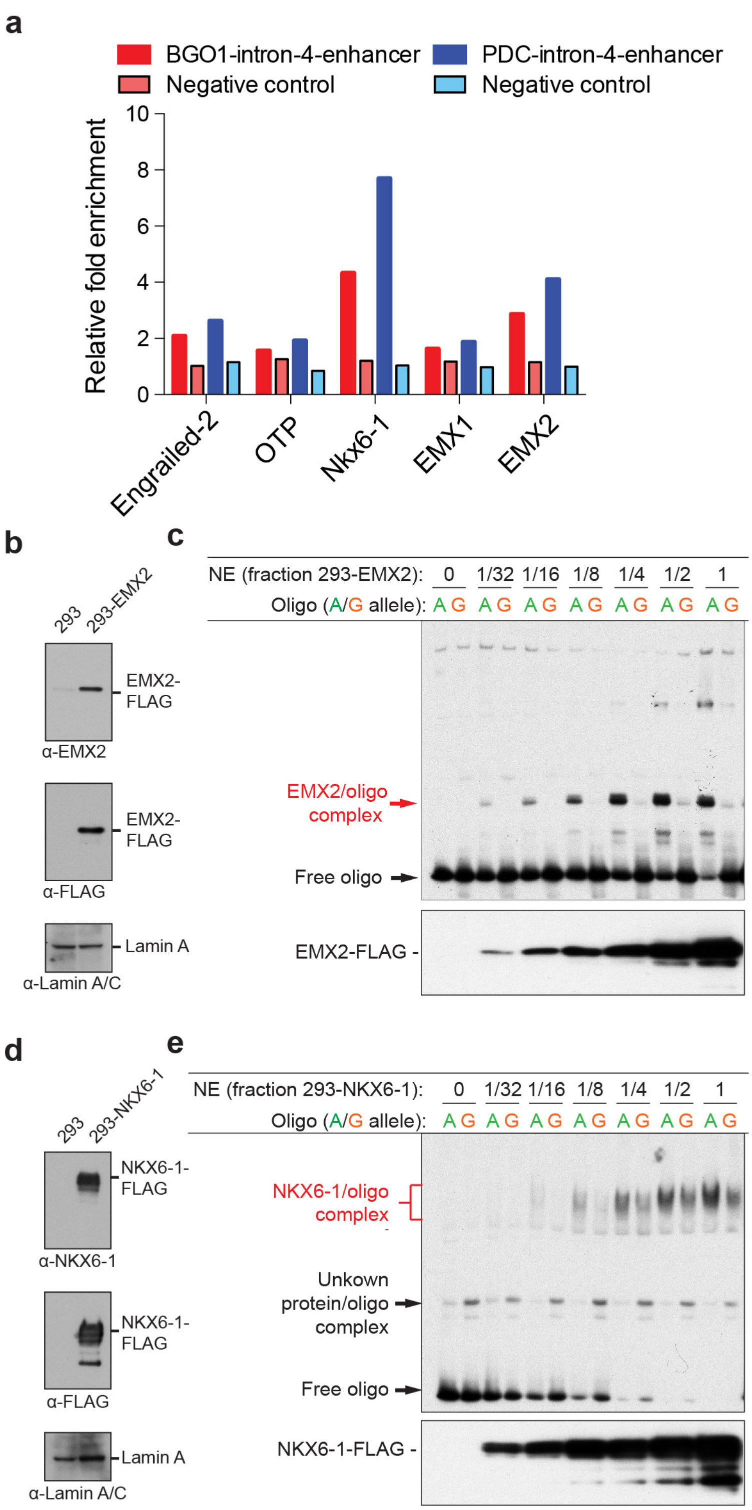
**b**



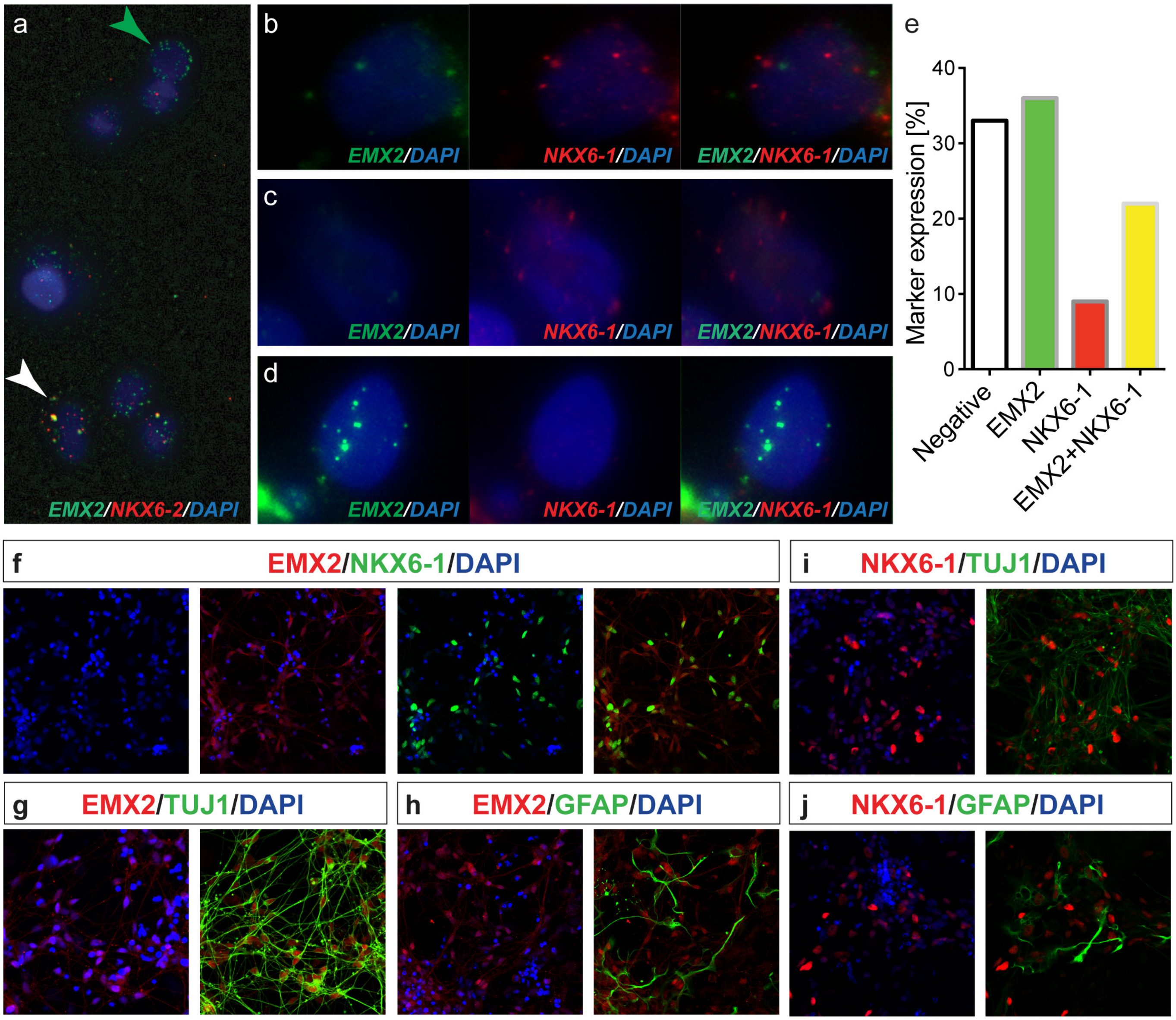

**Extended Data Figure 6 | Conditional GWAS and eQTL analysis to assess the effect of Parkinson's disease risk SNP rs356168. a**, Table showing baseline and conditional GWAS analysis in 5 publicly available PD GWAS cohorts (totalling 6,014 cases and 9,119 controls) to assess the extent to which rs356168 explains the observed associations in the *SNCA* locus of the top reported GWAS SNP rs356182 and the independent 5′ region SNP rs7681154 (ref. 12). **b**, qRT–PCR analysis for *SNCA* expression in postmortem frontal cortex tissue obtained from Parkinson's disease patients and controls stratified by risk genotype at rs356168 (number of brain samples from Parkinson's disease patients and controls are indicated for each genotype). Expression models were analysed including adjustment for disease status, sex, pH, age at death, as well as for the interaction between PMI and disease status. Significance was assessed using a one-sided test based on the *a priori* hypothesis of an association between the G allele at rs356168 and increased expression of *SNCA*. *P* values comparing each genotype are displayed in the graph. Alternatively analysis using linear regression shows a significant increase of total *SNCA* levels in carriers of the G allele at rs356168 ($P = 0.031$).



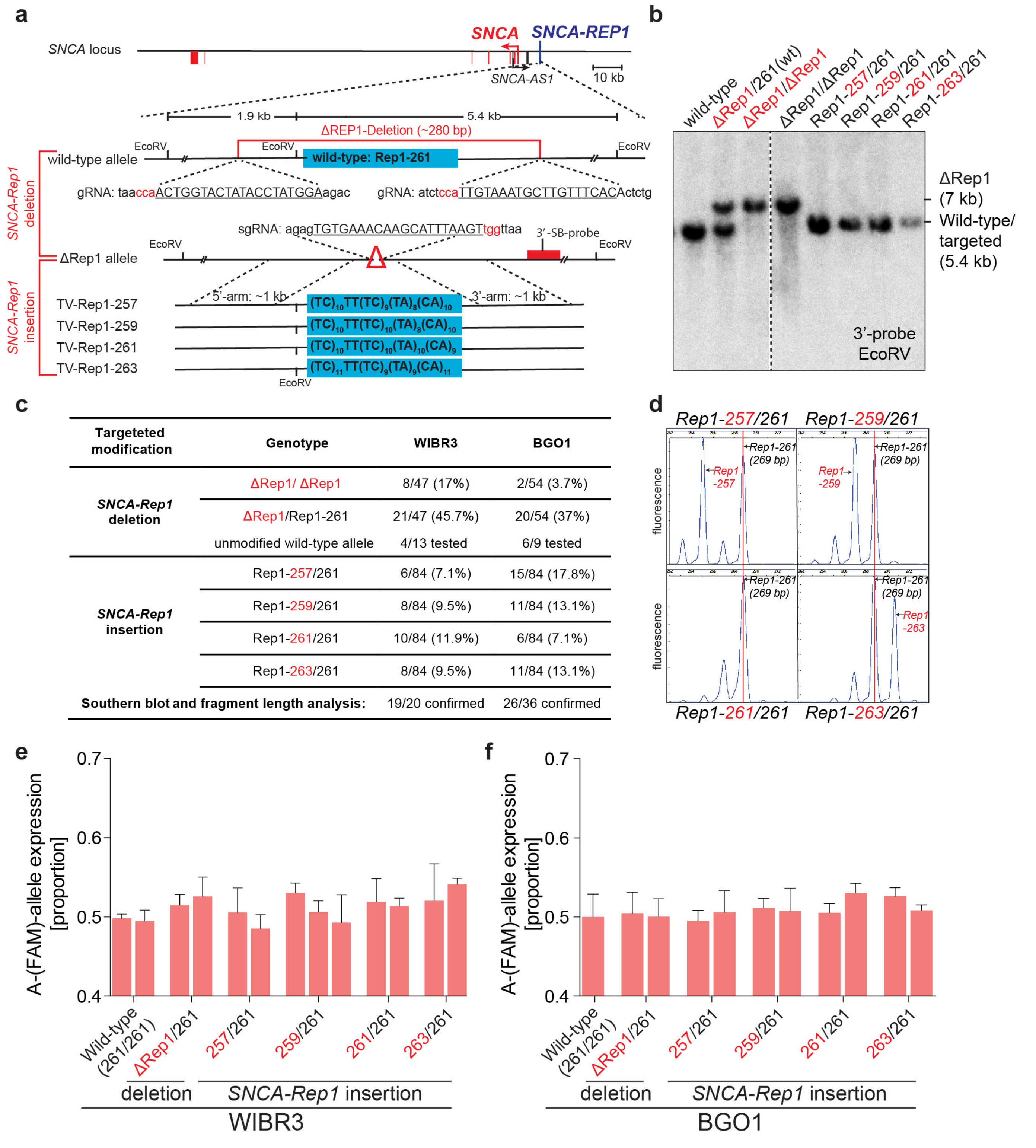

**Extended Data Figure 7 | Parkinson's disease associated risk variants have little effect on enhancer-specific chromatin modifications at the intron-4 enhancer. a**, Overview of isogenic cell lines (derived from WIBR3 human ES cells) carrying distinct genotypes of the intron-4 enhancer element used for chromatin-immunoprecipitation (ChIP-qRT–PCR and ChIP-seq). **b**, **c**, ChIP-qRT–PCR analysis in neurons derived from isogenic cell lines with indicated genotypes for binding of the enhancer-specific chromatin marks H3K4me1 (**b**) and H3K27ac (**c**) at the intron-4 and 3′ UTR enhancer sequences compared with indicated negative control regions (calculated as percent of input). **d**, Gene tracks of ChIP-seq analysis for the active enhancer mark H3K27ac in *in vitro* differentiated neurons derived from isogenic cell lines with indicated genotypes. **e**, Quantitative read density analysis of H3K27ac ChIP-seq data (as shown in **d**) displaying relative read density of the intron-4 enhancer compared to the 3′ UTR enhancer to control for variability between ChIP experiments. Source Data for this figure are available online.




**Extended Data Figure 8 | Sequence-specific binding of brain-expressed TFs EMX2 and NKX6-1 at *SNCA* intron-4 enhancer. a**, ChIP-qRT–PCR for binding of indicated TFs at the intron-4 enhancer element compared with negative control region in the *SNCA* locus (calculated as fold enrichment compared with IgG Isotype control) in human ES cell line BGO1 and human iPS cell line IPS-PDC[16] (derived from fibroblast AG20446). **b**, **d**, Western blot analysis in nuclear extracts (NE) (used for experiments displayed in Fig. 3b, c and Extended Data Fig. 8c, e) from HEK293 cells overexpressing Myc-DDK-tagged EMX2 (**b**) or NKX6-1 (**d**) using indicated antibodies against DDK-(Flag)-tag, EMX2 and NKX6-1 respectively. Lamin A was used to control for equal loading in NEs. **c**, **e**, EMSA analysis to determine TF concentration-dependent sequence-specific binding of EMX2 (**c**) and NKX6-1 (**e**) to oligonucleotides harbouring the indicated genotype at rs356168 (A/G allele). NEs from HEK293 cells overexpressing Myc-DDK(Flag)-tagged EMX2 (**c**) or NKX6-1 (**e**) were diluted with NEs from wild-type cells at indicated fractions to generate a TF concentration gradient. Relative EMX2 and NKX6-1 protein concentration in mixed samples were determined by western blot analysis using an antibody against the DDK(Flag)-tag (panel below respective EMSA). Red arrows point to oligonucleotide-specific binding of overexpressed TFs. Source Data for this figure are available online.




**Extended Data Figure 9 | Single-molecule mRNA FISH analysis and immunostaining for TFs EMX2 and NKX6-1 in mixed neuronal cultures. a**, Single-molecule mRNA FISH for EMX2-(Cy5) and NKX6-1-(Alexa594) (displayed in false colour) labelling EMX2 and NKX6-1 mRNA transcripts in WIBR3-derived *in vitro* differentiated neurons (differentiation day 21). Cultured neurons were dissociated before hybridization and attached to a glass slide before imaging. Representative image shows multiple cells, which are either single-positive for EMX2 (green arrowhead) or double-positive for EMX2 and NKX6-1 (white arrowhead). **b**–**d**, Representative images showing individual cells which are either double positive for the expression of EMX2 and NKX6-1 (**b**) or single positive for either NKX6-1 (**c**) or EMX2 (**d**). **e**, Quantification of 100 individual cells for the presence of EMX2 and NKX6-1 transcripts. **f**–**j**, Co-immunostaining for EMX2, NKX6-1, neuronal-specific beta-III-tubulin (TUJ1) and astrocyte-specific glial fibrillary acidic protein (GFAP) in mixed neuronal cultures. Source Data for this figure are available online.




**Extended Data Figure 10** | See next page for caption.

**Extended Data Figure 10 | *SNCA-Rep1* repeat length has no *cis*-acting effect on *SNCA* expression in human ES-cell-derived neurons.**
**a**, Schematic illustration of the CRISPR/Cas9-mediated genome editing strategy to generate heterozygous deletions (ΔRep1/wild-type) in WIBR3 and BGO1 human ES cells and subsequently insert indicated *SNCA-Rep1* length variants. Displayed are the genomic organization of the *SNCA* locus, an enlarged view of the wild-type and *SNCA-Rep1* deleted (ΔRep1) allele, gRNA-targeting sequences (underlined, PAM sequence in red), restriction sites and Southern blot (SB) probe. Shown below is targeting vector (TV) design to insert the respective *SNCA-Rep1* elements (*Rep1-257*, *Rep1-259*, *Rep1-261* and *Rep1-263*) with indicated repeat sequences. Only clones with heterozygous deletion of the repeat element on the same chromosome were identified based on the genotype of two heterozygous SNPs (rs58864428 and rs10030935) upstream and downstream of *SNCA-Rep1* and selected for subsequent experiments. **b**, Representative Southern blot analysis of wild type and targeted WIBR3 human ES cells with indicated *SNCA-Rep1* genotypes (modified alleles highlighted in red; unmodified wild-type allele represents *SNCA-Rep1-261* element). **c**, Table summarizing *SNCA-Rep1* deletions and insertions in WIBR3 and BGO1 human ES cells. **d**, Representative fragment length analysis confirms expected *SNCA-Rep-1* repeat length in targeted cell lines. Red line indicates *SNCA-Rep1-261* peak at 269 bp. **e**, **f**, Analysis of relative allele-specific *SNCA* expression in neurons (differentiation day 25) derived from targeted cells lines with indicated *SNCA-Rep1* alleles compared with untargeted controls in WIBR3 (**e**) and BGO1 (**f**) human ES cells (expression was normalized relative to wild-type cells). Shown are mean values ± s.d. of three independent biological replicate experiments for each individual clone of the indicated genotype. Differences between individual clones or combined clones by genotypes were not significant based on one-way ANOVA testing for multiple comparisons between groups (alpha = 0.05) and did not show a significant repeat length-dependent linear trend as analysed by linear regression. Source Data for this figure are available online.

LETTER

doi:10.1038/nature17949

# Unique human immune signature of Ebola virus disease in Guinea

Paula Ruibal[1,2,3,4]*, Lisa Oestereich[2,3,4]*, Anja Lüdtke[1,2,3,4]*, Beate Becker-Ziaja[2,3,4]*, David M. Wozniak[2,3,4], Romy Kerber[2,3,4], Miša Korva[4,5], Mar Cabeza-Cabrerizo[2,4], Joseph A. Bore[4], Fara Raymond Koundouno[4], Sophie Duraffour[2,4], Romy Weller[4,6], Anja Thorenz[4,7], Eleonora Cimini[4,8], Domenico Viola[4,8], Chiara Agrati[4,8], Johanna Repits[4], Babak Afrough[4,9], Lauren A. Cowley[4,10], Didier Ngabo[4,9], Julia Hinzmann[4,11], Marc Mertens[4,12], Inês Vitoriano[4,9], Christopher H. Logue[4,9], Jan Peter Boettcher[4,11], Elisa Pallasch[2,3,4], Andreas Sachse[4,11], Amadou Bah[4,13], Katja Nitzsche[3,4], Eeva Kuisma[4,9], Janine Michel[4,11], Tobias Holm[2,3,4], Elsa-Gayle Zekeng[4], Isabel García-Dorival[4,14], Roman Wölfel[3,4,15], Kilian Stoecker[3,4,15], Erna Fleischmann[3,4,15], Thomas Strecker[3,4,16], Antonino Di Caro[4,8], Tatjana Avšič-Županc[4,5], Andreas Kurth[4,11], Silvia Meschi[4,8], Stephane Mély[4,17], Edmund Newman[4,9], Anne Bocquin[4,17], Zoltan Kis[4,18,19], Anne Kelterbaum[3,4,16], Peter Molkenthin[3,4,15], Fabrizio Carletti[4,8], Jasmine Portmann[4,20], Svenja Wolff[3,4,16], Concetta Castilletti[4,8], Gordian Schudt[3,4,16], Alexandra Fizet[4,21], Lisa J. Ottowell[4,9], Eva Herker[1], Thomas Jacobs[2], Birte Kretschmer[22], Ettore Severi[19], Nobila Ouedraogo[11], Mar Lago[23], Anabel Negredo[24], Leticia Franco[24], Pedro Anda[24], Stefan Schmiedel[25], Benno Kreuels[2,3,25], Dominic Wichmann[3,25], Marylyn M. Addo[3,25], Ansgar W. Lohse[3,25], Hilde De Clerck[26], Carolina Nanclares[26], Sylvie Jonckheere[26], Michel Van Herp[26], Armand Sprecher[26], Gao Xiaojiang[27,28], Mary Carrington[27,28], Osvaldo Miranda[29], Carlos M. Castro[29], Martin Gabriel[2,3,4], Patrick Drury[30], Pierre Formenty[30], Boubacar Diallo[30], Lamine Koivogui[31], N'Faly Magassouba[32], Miles W. Carroll[4,9], Stephan Günther[2,3,4]§ & César Muñoz-Fontela[1,2,3,4]§

**Despite the magnitude of the Ebola virus disease (EVD) outbreak in West Africa, there is still a fundamental lack of knowledge about the pathophysiology of EVD[1]. In particular, very little is known about human immune responses to Ebola virus[2,3]. Here we evaluate the physiology of the human T cell immune response in EVD patients at the time of admission to the Ebola Treatment Center in Guinea, and longitudinally until discharge or death. Through the use of multiparametric flow cytometry established by the European Mobile Laboratory in the field, we identify an immune signature that is unique in EVD fatalities. Fatal EVD was characterized by a high percentage of $CD4^+$ and $CD8^+$ T cells expressing the inhibitory molecules CTLA-4 and PD-1, which correlated with elevated inflammatory markers and high virus load. Conversely, surviving individuals showed significantly lower expression of CTLA-4 and PD-1 as well as lower inflammation, despite comparable overall T cell activation. Concomitant with virus clearance, survivors mounted a robust Ebola-virus-specific T cell response. Our findings suggest that dysregulation of the T cell response is a key component of EVD pathophysiology.**

During the initial months of the EVD outbreak in Guinea, we transferred leftover blood samples from EVD patients diagnosed by the EMLab in Guéckédou ($n=47$) to Europe (Extended Data Fig. 1a and 1b). Immunophenotyping analysis of these samples, indicated a significantly higher expression of CTLA-4 in $CD8^+$ T cells from EVD patients compared to non-EVD patients ($n=61$) (Extended Data Fig. 1c). The levels of CTLA-4 were significantly higher in $CD8^+$ T cells from fatal EVD cases compared to survivors (Extended Data Fig. 1d). CTLA-4 plays an important role in inhibiting T cell function, an immune homeostasis mechanism to control excessive or persistent T cell activation[4,5]. Owing to its regulatory properties, there are licensed therapeutics to either antagonize or enforce CTLA-4 function[6,7]. Thus, we hypothesized that our findings could reflect a pathophysiological mechanism of EVD that might be amenable to therapeutics. However, we had concerns regarding the quality of the material that arrived at our laboratory after days of transport and we therefore established flow cytometry directly in Guinea to further evaluate T cell immunity in EVD. Leftover diagnostic blood samples from 157 EVD patients tested by the EMLab at the Coyah Ebola Treatment Center (ETC) were transferred to our laboratory in Conakry within 24 h after collection (Extended Data Fig. 2a, b). The median day of admission at the ETC for both fatalities and survivors was day 4 post symptom onset and, thus, outcome was not associated with the time elapsed between onset of disease and admission (Fig. 1a).

First, we evaluated the immune status of EVD patients upon their arrival to the ETC. Owing to the limitation to six parameters in our field analyses (Extended Data Fig. 3), we focused on the evaluation of individual markers of T cell function. We determined the expression levels of HLA-DR and Ki-67 as markers of activated T cells[8], as well as PD-1 and CTLA-4, which together are key regulators of T cell homeostasis[5,9]. The percentage of $CD4^+$ and $CD8^+$ T cells expressing

[1]Heinrich Pette Institute, Leibniz Institute for Experimental Virology, 20251 Hamburg, Germany. [2]Bernhard Nocht Institute for Tropical Medicine, World Health Organization Collaborating Center for Arbovirus and Hemorrhagic Fever Reference and Research, 20359 Hamburg, Germany. [3]German Center for Infection Research (DZIF), Partner Sites Hamburg, Munich, and Marburg, Germany. [4]European Mobile Laboratory Consortium, Bernhard-Nocht-Institute for Tropical Medicine, D-20359 Hamburg, Germany. [5]Institute of Microbiology and Immunology, Faculty of Medicine, University of Ljubljana, 1000 Ljubljana, Slovenia. [6]Institute of Experimental Virology, Twincore, Center for Experimental and Clinical Infection Research, 30625 Hannover, Germany. [7]Hannover Medical School, 30625 Hannover, Germany. [8]National Institute for Infectious Diseases 'Lazzaro Spallanzani', 00149 Rome, Italy. [9]Public Health England, Porton Down, Salisbury SP4 0JG, UK. [10]Public Health England, Colindale Ave, London NW9 5EQ, UK. [11]Robert Koch Institute, 13353 Berlin, Germany. [12]Friedrich Loeffler Institute, 17493 Greifswald-Island of Riems, Germany. [13]Swiss Tropical and Public Health Institute, 4051 Basel, Switzerland. [14]Institute of Infection and Global Health, University of Liverpool, Liverpool L69 7BE, UK. [15]Bundeswehr Institute of Microbiology, 80937 Munich, Germany. [16]Institute of Virology, Philipps University, 35043 Marburg, Germany. [17]Laboratoire P4-Jean Mérieux, US003 INSERM, 69365 Lyon, France. [18]National Center for Epidemiology, Hungarian National Biosafety Laboratory, H1097 Budapest, Hungary. [19]European Centre for Disease Prevention and Control, 171 65 Solna, Sweden. [20]Federal Office for Civil Protection, CH-3700 Spiez, Switzerland. [21]Unité de Biologie des Infections Virales Emergentes, Institut Pasteur, 69365 Lyon, France. [22]Eurice, European Research and Project Office, 10115 Berlin, Germany. [23]Infectious Diseases Unit, Internal Medicine Service, Hospital La Paz, 28046 Madrid, Spain. [24]National Center of Microbiology, Institute of Health 'Carlos III', 28220 Madrid, Spain. [25]University Medical Center Hamburg-Eppendorf, 20246 Hamburg, Germany. [26]Médecins sans Frontières, B-1050 Brussels, Belgium. [27]Cancer and Inflammation Program, Laboratory of Experimental Immunology, Leidos Biomedical Research, Frederick National Laboratory for Cancer Research, Frederick, Maryland 21702, USA. [28]Ragon Institute of MGH, MIT and Harvard, Cambridge, Massachusetts 02139, USA. [29]Hospital Militar Central Dr. Carlos J. Finlay, 11400 Havana, Cuba. [30]World Health Organization, 1211 Geneva 27, Switzerland. [31]Institut National de Santé Publique, 2101 Conakry, Guinea. [32]Université Gamal Abdel Nasser de Conakry, CHU Donka, 2101 Conakry, Guinea.
*These authors contributed equally to this work.
§These authors jointly supervised this work.

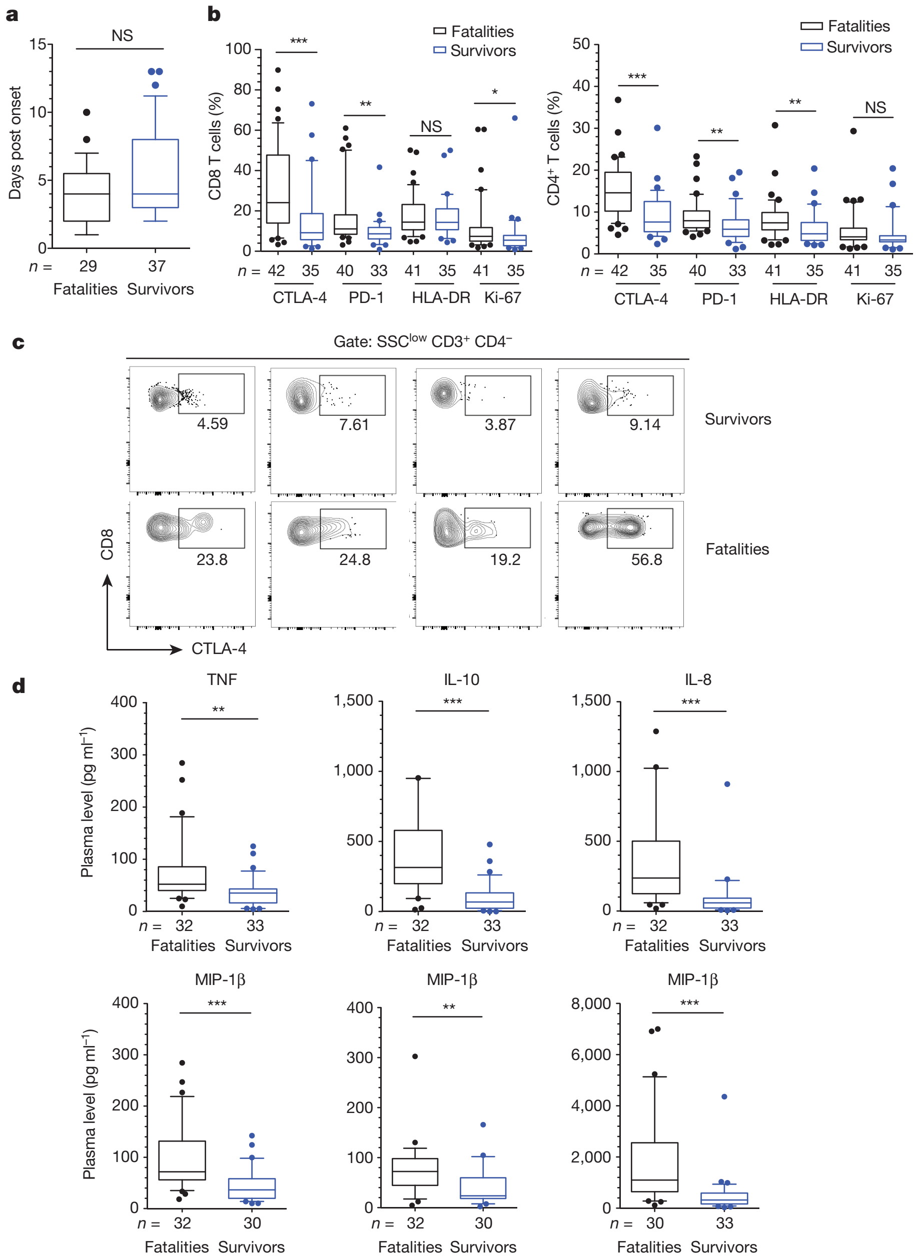



**Figure 1 | Immune signature of EVD at the time of admission. a**, Day of symptom onset at admission. **b**, Frequencies of CD8$^+$ T cells (left) and CD4$^+$ T cells (right) positive for the indicated markers. Black boxes represent fatal cases and blue boxes represent survivors. **c**, Plots of four survivors and four fatalities showing frequencies of CTLA-4$^+$ CD8$^+$ T cells. **d**, Levels of plasma cytokines in fatal versus surviving EVD cases. Statistical analysis was performed by Mann–Whitney test: NS, not significant; $*P \leq 0.05$; $**P \leq 0.01$; $***P \leq 0.001$.

CTLA-4 and PD-1 was significantly higher in fatal cases compared with survivors (Fig. 1b). Sometimes more than 50% of total CD8$^+$ T cells expressed CTLA-4 (Fig. 1c). Fatal cases also showed higher frequency of Ki-67$^+$ CD8$^+$ T cells and HLA-DR$^+$ CD4$^+$ T cells than survivors (Fig. 1b).

High expression of both PD-1 and CTLA-4 has been correlated with functional exhaustion (loss of function) of T cells in chronic inflammatory conditions[9–11]. However, the role of these molecules during acute infection is less well understood. In previous studies, expression of CTLA-4 and PD-1 in T cells correlated with the extent of pro-inflammatory responses against hantavirus and influenza virus among others[12–14]. Thus, we next evaluated the levels of pro- and anti-inflammatory cytokines in fatal and non-fatal EVD. In agreement with previous reports[15,16], fatal EVD cases had significantly higher levels of serum pro-inflammatory cytokines (TNF (also known as TNFα) and IL-8) (Fig. 1d). Moreover, the levels of pro-inflammatory chemokines (MIP-1α, MIP-1β and MCP1) were also significantly upregulated in fatalities (Fig. 1d). Levels of anti-inflammatory cytokines such as IL-10 were also higher in fatalities suggesting the onset of compensatory homeostatic mechanisms in response to excessive inflammation, consistent with expression of T cell inhibitory molecules such as CTLA-4 and PD-1.

To gain insight into the correlation between CTLA-4 expression and the functional status of T cells in patients we performed analyses of T cells from cryopreserved peripheral blood leukocytes, which allowed evaluation of T cell marker co-expression (Extended Data Fig. 4). We used co-expression of CD38 and HLA-DR as well as CD38 and Ki-67 to define activated T cells[3,8,13], and co-expression of CTLA-4 and PD-1 to identify effector T cells that had activated inhibitory mechanisms[13]. Our results indicated robust CD8$^+$ T cell activation in both fatal and non-fatal EVD cases with no statistically significant difference between the two groups. The frequency of activated CD8$^+$ T cells (around 30% of CD38$^+$ HLA-DR$^+$ and 18% of CD38$^+$ Ki-67$^+$) was similar to that previously reported in acute surviving medevac EVD patients[3], and to that described in other acute infections[13] and after vaccination[8]. However, the frequency of CD8$^+$ T cells co-expressing PD-1 and CTLA-4 was significantly higher in fatalities (Fig. 2a, b). The results with CD4$^+$ T cells corresponded to those of CD8$^+$ T cells, with no differences in the activation status between fatal and non-fatal EVD, but a higher frequency of cells co-expressed CTLA-4 and PD-1 in fatal

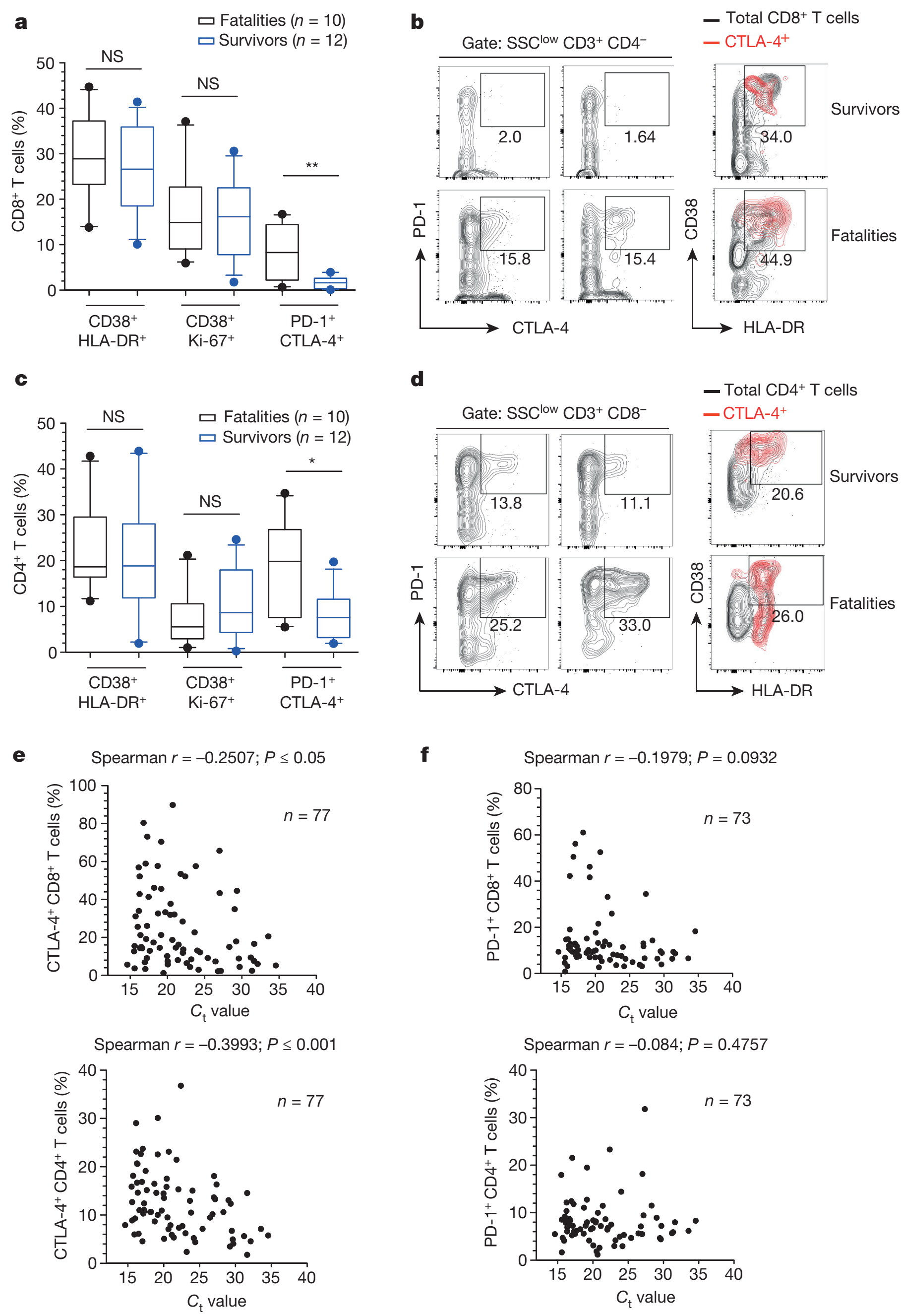



**Figure 2 | Functional properties of T cells in EVD patients. a**, Frequencies of CD8$^+$ T cells co-expressing the indicated markers. **b**, Plots of two surviving and two fatal EVD patients depicting CD8$^+$ T cell populations co-expressing the indicated markers. Overlays indicate activation status of CTLA-4$^+$ CD8$^+$ T cells (red). Total CD8$^+$ T cells are represented by black contours. **c**, CD4$^+$ T cells co-expressing the indicated markers. **d**, Plots of two surviving and two fatal patients depicting CD4$^+$ T cell populations co-expressing the indicated markers. Overlays indicate activation status of CTLA-4$^+$ CD4$^+$ T cells (red). Total CD4$^+$ T cells are represented by black contours. **e**, Correlation between frequency of CD8$^+$ T cells (upper graph) and CD4$^+$ T cells (lower graph) expressing CTLA-4 and $C_t$ values. **f**, Correlation between the frequency of CD8$^+$ T cells (upper graph) and CD4$^+$ T cells (lower graph) positive for PD-1 and $C_t$ values. Statistic analysis was performed by Mann–Whitney test: NS, not significant; $*P \leq 0.05$; $**P \leq 0.01$.

cases (Fig. 2c, d). CTLA-4$^+$ T cells in both CD8$^+$ and CD4$^+$ subsets expressed high levels of CD38 and HLA-DR, suggesting a discrete subset of effector T cells that have initiated a compensatory homeostatic mechanism (Fig. 2b, d).

Whether or not CTLA-4 overexpression affects T cell-mediated viral clearance is controversial. While some reports have correlated the frequency of CD8$^+$ T cells expressing CTLA-4 with poor viral clearance[14], others do not support this hypothesis[17]. To explore the relationship between CTLA-4 and PD-1 expression and virus loads during EVD, we analysed whether the percentage of T cells expressing CTLA-4, PD-1, or both, correlated with the threshold cycle ($C_t$) values of the EBOV real-time PCR. Fatal EVD cases had significantly lower $C_t$ values at the time of admission in agreement with the $C_t$ value being a strong predictor of outcome[18–20] (Extended Data Fig. 2c–e). The $C_t$ value negatively correlated with the percentage of T cells expressing CTLA-4 as well as the percentage of CD8$^+$ T cells co-expressing CTLA-4 and PD-1 (Fig. 2e and Extended Data Fig. 5a). However, the percentage of T cells expressing PD-1 alone did not correlate with the $C_t$ value (Fig. 2f). In summary, expression of CTLA-4 alone and in combination with PD-1 correlated with high viraemia.

Acute lymphopenia triggers proliferation of naive T cells with very low expression of CTLA-4 and PD-1 (ref. 21). Thus, we reasoned that differences in Ebola virus (EBOV)-induced lymphopenia[2], could be a confounding variable in our study. However, there were no differences in circulating peripheral blood T cells between fatal and non-fatal EVD cases (Extended Data Fig. 5b). A recent study comparing infection with the Makona versus Mayinga variants of EBOV in non-human primates found evidence of initial leukocytosis followed by moderate lymphopenia with no overall differences between both EBOV variants[22]. Initial leukocytosis in patients infected with the Makona variant of EBOV has been also observed[23,24], which indicates the need of longitudinal evaluation of haematological parameters in patients infected with EBOV Makona, an approach that was not possible in our study due to the lack of statistically relevant numbers of fatal cases with longitudinal sampling.

To link the observed activation of CD8+ T cells with EBOV-specific responses, we sought to track EBOV-specific CD8+ T cells during infection through the use of HLA class I dextramers. We first performed *in silico* analysis of EBOV nucleoprotein-derived peptides predicted to bind with high affinity to selected HLA alleles common in West Africa (HLA-A*02:01, HLA-A*23:01 and HLA-B*35:01)[25] and designed dextramers (Extended Data Fig. 6). We chose the EBOV nucleoprotein based on previous reports indicating that nucleoprotein drives most of the CD8+ T cell response[26,27]. Dextramer-matching HLA alleles were identified in 26 patients via sequencing of the HLA locus. Six patients had more than one sample for longitudinal evaluation of EBOV-specific T cell responses. These included two fatal cases (F1 and F2) and four survivors (S2–S5). S3 and S4 were treated with favipiravir, and all the other patients received only supportive therapy. At the time of admission, the differences in the percentage of EBOV-specific T cells between fatalities and survivors were not statistically significant (Fig. 3a). However, differences became evident during later stages of infection. The fatal cases with longitudinal sampling (F1 and F2) showed a high frequency of PD-1+ CTLA-4+ CD8+ T cells until death, but barely detectable EBOV-specific CD8+ T cells (Fig. 3b, e). In contrast, an increase in the frequency of EBOV-specific CD8+ T cells was observed in survivors in coincidence with virus clearance and decrease of the PD-1+ CTLA-4+ CD8+ T cell population (Fig. 3c–e).

These findings were substantiated by the kinetics of CTLA-4 expression in two EVD patients, one fatality and one survivor, evacuated to Europe, as well as five survivors treated in Coyah, Guinea. Longitudinal analysis of these patients revealed persistent upregulation of CTLA-4 in the fatal case and transient upregulation in survivors (Extended Data Fig. 7). Of note, the increase of EBOV-specific T cells in the surviving patients coincided with contraction rather than expansion of the CD38+ HLA-DR+ subset (Fig. 3e). Despite the fact that the phenotype of activated T cells (CD38+ HLA-DR+) suggests engagement of the T cell receptor (TCR) and not bystander T cell activation[8], our findings may point to the presence of non-EBOV-specific T cells within the CD38+ HLA-DR+ compartment.

The main hypothesis we formulate on the basis of these data is that differences between the T cell response of fatal and surviving EVD patients are centred in the mechanisms that regulate T cell homeostasis. While all patients showed T cell activation irrespective of outcome, it was the expression of the regulatory molecules CTLA-4 and PD-1 on peripheral blood T cells that marked fatal EVD and correlated with high viraemia, a known predictor of poor outcome[19,20]. We hypothesize that this upregulation reflects a compensatory mechanism to excess inflammatory stimuli, which is consistent with the concomitant expression of pro-inflammatory cytokines in fatal cases. However, our findings do not indicate causality between expression

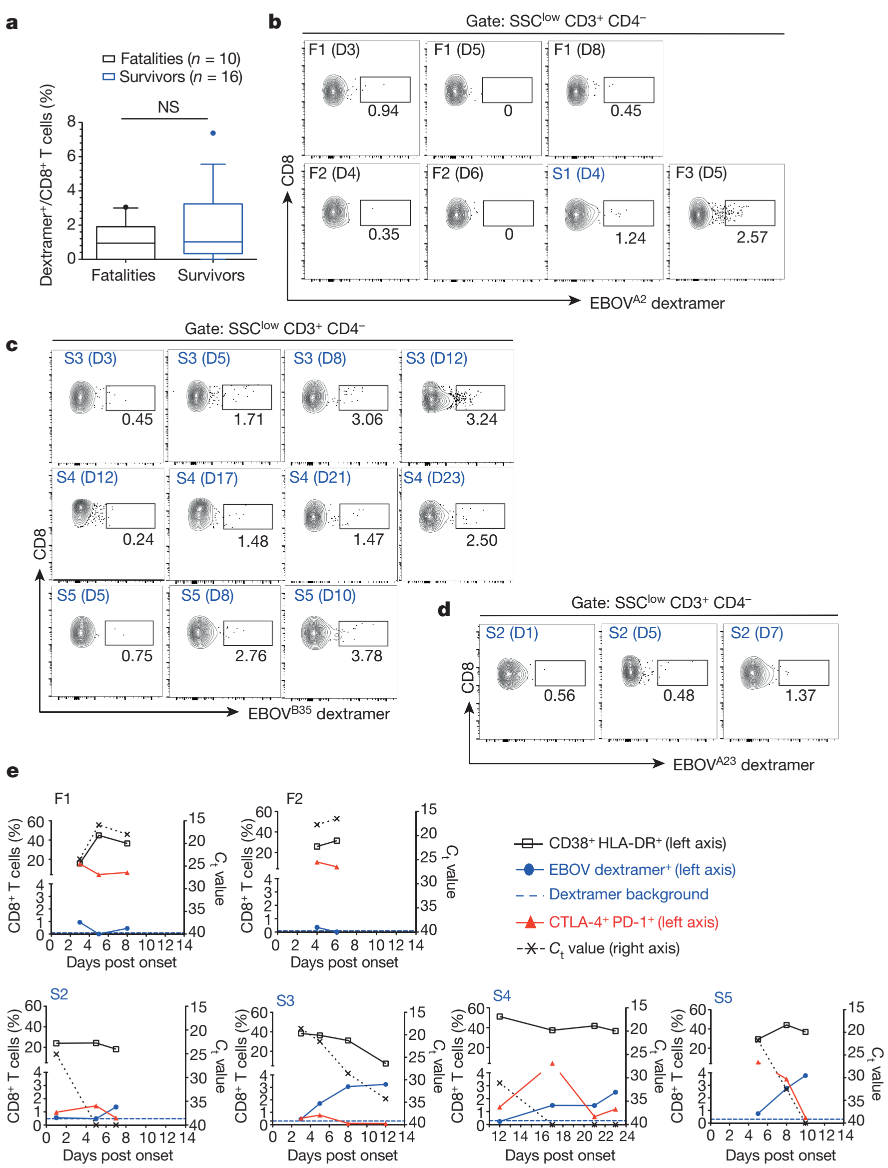


**Figure 3 | Longitudinal evaluation of EBOV-specific T cell immunity. a**, Frequency of EBOV-specific CD8+ T cells in EVD patients at admission. Statistic analysis was performed by Mann–Whitney test: NS, not significant. **b**, Frequency of CD8+ T cells specific for the HLA-A*02:01-restricted peptide FLSFASLFL. Two fatal patients (F1 and F2, see panel **d**) are shown and day of sample evaluation post onset is indicated in parentheses. A surviving patient (S1) and fatal patient (F3) are shown for positive HLA-A*02:01 dextramer staining. **c**, Frequencies of CD8+ T cells specific for the HLA-B*35:01-restricted FPQLSAIAL peptide are shown in survivors (S3, S4 and S5) with serial sampling. **d**, Plots of one survivor (S2) showing staining of EBOV-specific CD8+ T cells restricted for HLA-A*23:01 AYQGDYKLF. **e**, Longitudinal patient data at the indicated days after symptom onset. Left axis represents the frequency of CD8+ T cells co-expressing CD38 and HLA-DR (black), CTLA-4 and PD-1 (red) and dextramer-positive cells (blue). Dextramer background was assessed in HLA-matched healthy donors (Extended Data Fig. 6b), and background dextramer staining is shown as dashed blue lines. The corresponding flow cytometry plots are shown in panels **a**–**d**. $C_t$ values are represented for each patient with dashed lines and values depicted in the right axis.

of CTLA-4/PD-1 and poor outcome: on one hand it is possible that high viraemia triggers overwhelming inflammation thereby causing CTLA-4 and PD-1 upregulation on T cells, and on the other hand, it is plausible to assume that expression of these molecules inhibits T cell function leading to poor viral clearance. Indeed, both possibilities have been reported in the literature[13,28] and their elucidation requires evaluation of CTLA-4 function during EVD in adequate infection models. Interestingly, during severe EVD, both PD-1 and CTLA-4 are upregulated while in other acute human infections such as hantavirus disease and paediatric influenza, only CTLA-4, not PD-1, was upregulated[13,14]. These results suggest that different infections trigger specific regulatory mechanisms, the balance of which is vital for optimal T cell function. Robust T cell activation was observed in both fatal and non-fatal EVD cases as reflected by similar levels of discrete $CD8^+$ and $CD4^+$ T cells co-expressing CD38 and HLA-DR. These results are in agreement with the strong T cell activation observed in surviving US patients[3] and strengthen the notion that EVD is characterized by immune activation. Although limited, our dextramer data suggest that both fatal and non-fatal cases are able to mount EBOV-specific T cell responses. However, while in survivors the $PD\text{-}1^+$ $CTLA\text{-}4^+$ $CD8^+$ T cell compartment contracted in conjunction with viral clearance, formation of EBOV-specific T cells and recovery, this did not seem to be the case in fatal EVD. One plausible explanation for this finding is that the high expression of CTLA-4 observed in activated T cells of fatal cases promotes cell extrinsic inhibition of EBOV-specific T cells. T-cell-extrinsic CTLA-4 functions have been demonstrated *in vivo* and involve different mechanisms such as stripping of T cell co-stimulatory molecules from antigen-presenting cells, stimulation of regulatory T cells and local tryptophan starvation among others[5,29,30]. Further functional experiments with adequate *in vivo* models are needed to explore the role of regulatory T cell molecules during EVD and to test whether their modulation may serve as a putative post-exposure therapy against EVD.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 3 July 2015; accepted 1 April 2016.**

1. Messaoudi, I., Amarasinghe, G. K. & Basler, C. F. Filovirus pathogenesis and immune evasion: insights from Ebola virus and Marburg virus. *Nat. Rev. Microbiol.* **13,** 663–676 (2015).
2. Baize, S. *et al.* Defective humoral responses and extensive intravascular apoptosis are associated with fatal outcome in Ebola virus-infected patients. *Nat. Med.* **5,** 423–426 (1999).
3. McElroy, A. K. *et al.* Human Ebola virus infection results in substantial immune activation. *Proc. Natl Acad. Sci. USA* **112,** 4719–4724 (2015).
4. Best, J. A. *et al.* Transcriptional insights into the $CD8^+$ T cell response to infection and memory T cell formation. *Nat. Immunol.* **14,** 404–412 (2013).
5. Walker, L. S. K. & Sansom, D. M. The emerging role of CTLA4 as a cell-extrinsic regulator of T cell responses. *Nat. Rev. Immunol.* **11,** 852–863 (2011).
6. Linsley, P. S. *et al.* Immunosuppression *in vivo* by a soluble form of the CTLA-4 T cell activation molecule. *Science* **257,** 792–795 (1992).
7. Kvistborg, P. *et al.* Anti-CTLA-4 therapy broadens the melanoma-reactive $CD8^+$ T cell response. *Sci. Transl. Med.* **6,** 254ra128 (2014).
8. Miller, J. D. *et al.* Human effector and memory $CD8^+$ T cell responses to smallpox and yellow fever vaccines. *Immunity* **28,** 710–722 (2008).
9. Okazaki, T., Chikuma, S., Iwai, Y., Fagarasan, S. & Honjo, T. A rheostat for immune responses: the unique properties of PD-1 and their advantages for clinical application. *Nat. Immunol.* **14,** 1212–1218 (2013).
10. Kaufmann, D. E. *et al.* Upregulation of CTLA-4 by HIV-specific $CD4^+$ T cells correlates with disease progression and defines a reversible immune dysfunction. *Nat. Immunol.* **8,** 1246–1254 (2007).
11. Wherry, E. J. T cell exhaustion. *Nat. Immunol.* **12,** 492–499 (2011).
12. Hafalla, J. C. R. *et al.* The CTLA-4 and PD-1/PD-L1 inhibitory pathways independently regulate host resistance to plasmodium-induced acute immune pathology. *PLoS Pathog.* **8,** e1002504 (2012).
13. Lindgren, T. *et al.* Longitudinal analysis of the human T cell response during acute hantavirus infection. *J. Virol.* **85,** 10252–10260 (2011).
14. Ayukawa, H. *et al.* Expression of CTLA-4 (CD152) in peripheral blood T cells of children with influenza virus infection including encephalopathy in comparison with respiratory syncytial virus infection. *Clin. Exp. Immunol.* **137,** 151–155 (2004).
15. Hutchinson, K. L. & Rollin, P. E. Cytokine and chemokine expression in humans infected with Sudan Ebola virus. *J. Infect. Dis.* **196** (Suppl 2), S357–S363 (2007).
16. McElroy, A. K. *et al.* Ebola hemorrhagic fever: novel biomarker correlates of clinical outcome. *J. Infect. Dis.* **210,** 558–566 (2014).
17. Raué, H.-P. & Slifka, M. K. Pivotal advance: CTLA-4+T cells exhibit normal antiviral functions during acute viral infection. *J. Leukoc. Biol.* **81,** 1165–1175 (2007).
18. Schieffelin, J. S. *et al.* Clinical illness and outcomes in patients with Ebola in Sierra Leone. *N. Engl. J. Med.* **371,** 2092–2100 (2014).
19. Lanini, S. *et al.* Blood kinetics of Ebola virus in survivors and nonsurvivors. *J. Clin. Invest.* **125,** 4692–4698 (2015).
20. de La Vega, M. A. *et al.* Ebola viral load at diagnosis associates with patient outcome and outbreak evolution. *J. Clin. Invest.* **125,** 4421–4428 (2015).
21. Shvets, A. *et al.* Impaired negative regulation of homeostatically proliferating T cells. *Blood* **113,** 622–625 (2009).
22. Marzi, A. *et al.* Delayed disease progression in cynomolgus macaques infected with Ebola virus Makona strain. *Emerg. Infect. Dis.* **21,** 1777–1783 (2015).
23. Kreuels, B. *et al.* A case of severe Ebola virus infection complicated by Gram-negative septicemia. *N. Engl. J. Med.* **371,** 2394–2401 (2014).
24. Wolf, T. *et al.* Severe Ebola virus disease with vascular leakage and multiorgan failure: treatment of a patient in intensive care. *Lancet* **385,** 1428–1435 (2015).
25. Modiano, D. *et al.* HLA class I in three West African ethnic groups: genetic distances from sub-Saharan and Caucasoid populations. *Tissue Antigens* **57,** 128–137 (2001).
26. Sundar, K., Boesen, A. & Coico, R. Computational prediction and identification of HLA-A2.1-specific Ebola virus CTL epitopes. *Virology* **360,** 257–263 (2007).
27. Simmons, G. *et al.* Identification of murine T-cell epitopes in Ebola virus nucleoprotein. *Virology* **318,** 224–230 (2004).
28. de la Fuente, H., Cibrián, D. & Sánchez-Madrid, F. Immunoregulatory molecules are master regulators of inflammation during the immune response. *FEBS Lett.* **586,** 2897–2905 (2012).
29. Qureshi, O. S. *et al.* Trans-endocytosis of CD80 and CD86: a molecular basis for the cell-extrinsic function of CTLA-4. *Science* **332,** 600–603 (2011).
30. Corse, E. & Allison, J. P. Cutting edge: CTLA-4 on effector T cells inhibits in trans. *J. Immunol.* **189,** 1123–1127 (2012).

**Acknowledgements** We thank the African Union, MSF and WHO field teams and the Guinean health authorities in Guéckédou, Coyah and Conakry for their commitment and cooperation. We also thank J. Snyder-Cappione (Boston University) and M. Altfeld (Heinrich Pette Institute) for technical support. The EMLab is a technical partner of the WHO Emerging and Dangerous Pathogens Laboratory Network (EDPLN), and the Global Outbreak Alert and Response Network (GOARN) and the deployments in West Africa have been coordinated and supported by the GOARN Operational Support Team at WHO/HQ. This work was carried out in the context of the project EVIDENT (Ebola virus disease: correlates of protection, determinants of outcome, and clinical management) that received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No. 666100 and in the context of service contract IFS/2011/272-372 funded by Directorate-General for International Cooperation and Development. The project was further supported by grant GU 883/4-1 from the German Research Foundation. This project has been funded in in part with funds from the Spanish National Plan for Research and Development ISCIII and FEDER RD12/0018/006 as well as federal funds from the Frederick National Laboratory for Cancer Research, under Contract No. HHSN261200800001E. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the US Government. This research was supported in part by the Intramural Research Program of the NIH, Frederick National Laboratory, Center for Cancer Research. A.L. is a recipient of a pre-doctoral fellowship from the Leibniz Center of Infection. Z.K. is a recipient of a fellowship from the European Program for Public Health Microbiology Training (EUPHEM).

**Author Contributions** P.R., L.O., A.L. and B.B.-Z. performed all the experiments of this study and contributed equally. Authors affiliated with the Cuban medical teams, Médecins sans Frontières, Infectious Diseases Unit, Internal Medicine Service, Hospital La Paz, and the University Medical Center Hamburg-Eppendorf treated all the patients in this study and provided samples. All authors affiliated with the European Mobile Laboratory Consortium performed EVD diagnostics, collected samples in the field and/or coordinated EMLab operations and sample shipment to Conakry and Hamburg. Authors affiliated with WHO provided the logistics necessary to perform this study. A.S., E.S., P.F., N.O. and B.D. provided patient outcome, epidemiology and demographic data. E.H. and T.J. provided support for data analysis. B.K. provided logistic support for the EVIDENT project teams. G.X. and M.C. performed HLA typing. S.G. coordinated the study. C.M.-F. designed the study, performed experiments and wrote the manuscript with S.G. All authors discussed the results and commented on the manuscript.

**Author Information** Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to C.M.-F. (cesar.munoz-fontela@hpi.uni-hamburg.de) or S.G. (guenther@bni.uni-hamburg.de).

## METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

**Patients.** EVD patients included in the study were managed at the Ebola Treatment Centers in Guéckédou ($n = 47$) and Coyah ($n = 157$) under the medical care of Médecins sans Frontières and doctors deployed by the Cuban government, respectively. Patients with malaria co-infection were excluded from the study. Two patients evacuated into Europe for medical treatment were included in the study for longitudinal analysis. One patient was treated at the University Medical Center Hamburg-Eppendorf, Hamburg, Germany, and the other one at the Hospital La Paz, Madrid, Spain. Demographic and outcome data for EVD patients were obtained from databases of the World Health Organization. The National Committee of Ethics in Medical Research of Guinea as well as the Ethics Committee of the Medical Association of Hamburg approved the use of diagnostic leftover samples and corresponding patient data for this study (permits N°11/CNERS/14 and PV4910). This study was also approved by the protocol review office of the US National Cancer Institute institutional review board. As the samples had been collected as part of the public health response to contain the outbreak in Guinea, informed consent was not obtained from patients. Informed consent was obtained from the two EVD patients treated in Hamburg and Madrid.

**Study samples and flow cytometry analysis.** Real-time RT–PCR was performed on EDTA-blood of patients with suspected EVD using the RealStar Ebolavirus RT–PCR Kit 1.0 (Altona Diagnostics) at the European Mobile Laboratory (EMLab) units in Guéckédou and Coyah. Malaria was diagnosed using a rapid test. Whole blood samples from Guéckédou were shipped to the biosafety level 4 (BSL-4) laboratory in Hamburg within 1–3 weeks after collection and processed immediately upon arrival via multiparametric flow cytometry.

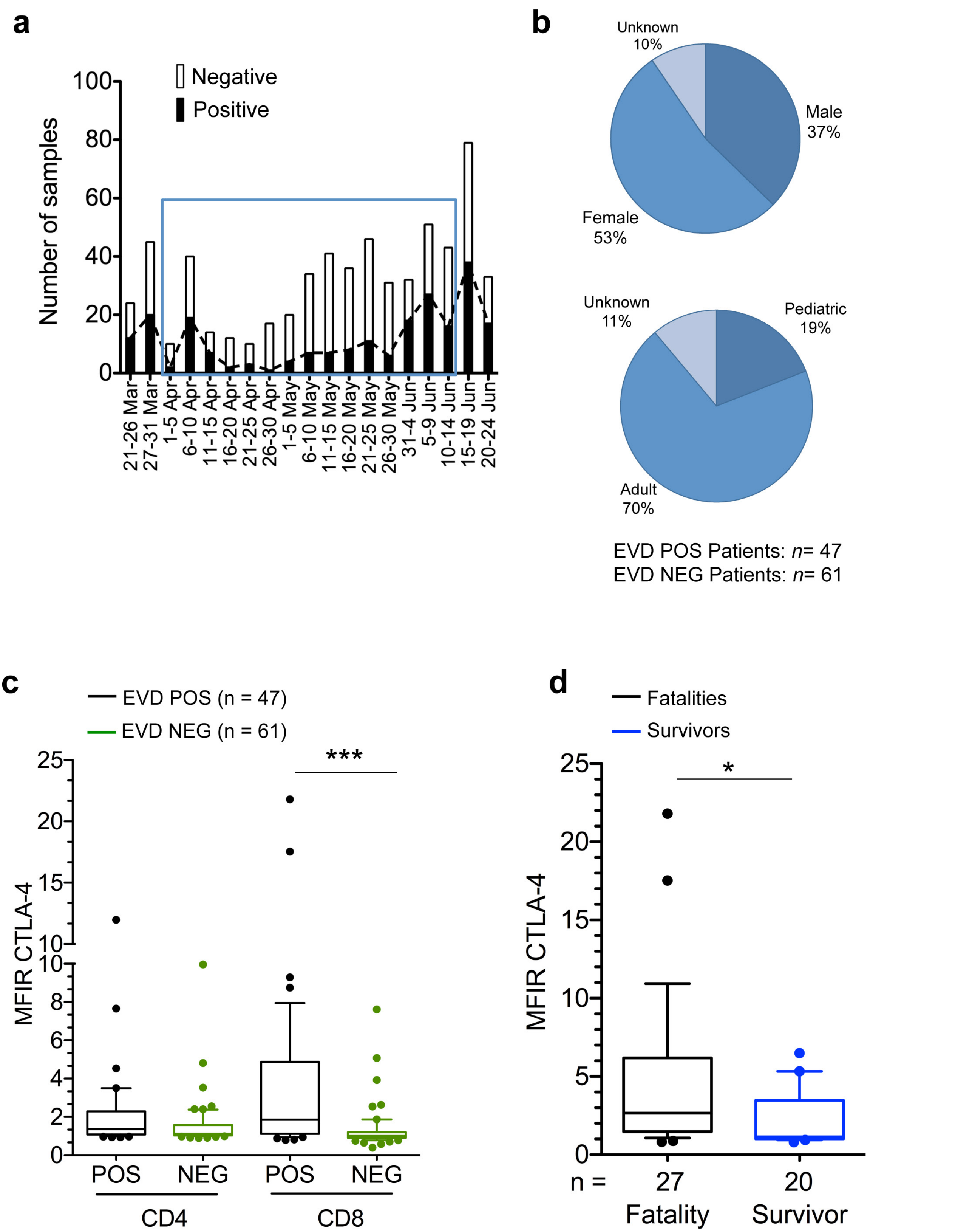
Leftover samples from Coyah were shipped within 24 h after collection to the EMLab immunology laboratory at Donka Hospital in Conakry. PBMCs were isolated after sedimentation of cells in EDTA whole blood tubes. Red blood cells were lysed with Red Blood Cell Lysing buffer (BD Biosciences). PBMCs were processed immediately upon reception for immunophenotypic analysis or were cryopreserved and transported to the BSL-4 laboratory in Hamburg. Immune phenotyping was achieved via multiparametric flow cytometry panel using a battery of commercially available antibodies as follows: anti-human (h) CD3-APCCy7 (clone UCHT1), anti-hCD4-FITC (clone OKT4), anti-hCD8-PeCy7 (clone RPAT8), anti-hCD152 (CTLA-4)-PE (clone L3D10), anti-hCD279 (PD-1)-APC (clone EH12-2HT), anti-hHLA-DR-PerCP Cy5.5 (clone L243), anti-hKi67-PE (clone 16A8), anti-hCD38-APC (clone HB-7). All antibodies were from Biolegend. Single-cell PBMC suspensions were incubated with live/dead discrimination dye (Zombie-NIR from Biolegend) followed by FACS block (Human TruStain Fc receptor blocking antibodies from Biolegend) for 20 min followed by staining with antibodies against extracellular antigens. After extracellular staining samples were inactivated in Cytofix/Cytoperm (BD) buffer in the presence of 4% formaldehyde followed by staining with intracellular antibodies (anti-CD3, Ki-67 and CTLA-4). Sample acquisition was done in a Guava easyCyte 8 Flow Cytometer from Millipore in Guinea. In Hamburg, samples were thawed at 37 °C and the sample volume transferred to ice-cold 15-ml falcon tubes where 5 ml ice-cold R8 medium (RPMI + 8% FBS) were added. Sample tubes were then centrifuged to remove the cryopreservant (10% DMSO) at 1,000 r.p.m. for 10 min and resuspended in R8 medium. Samples were centrifuged once more and the pelleted cells were washed with 12 ml of ice-cold R8 medium. Samples were processed as indicated above. In experiments involving dextramer staining, samples were incubated in a volume of 50 μl of PBS with 5 μl of the indicate dextramers (1:10 dilution) before extracellular antibody staining. All dextramers were purchased from Immudex. Samples were acquired in a LSR Fortessa instrument (BD). Flow cytometry analysis was done with FlowJO software (Treestar).

***In silico* peptide analysis.** Prediction of EBOV nucleoprotein-derived peptide binding to selected human HLA alleles was achieved using an artificial neural network method at the Immune Epitope Database and Analysis Resource (IEDB) (http://www.iedb.org). Selected peptides ($IC_{50} < 50$ nM) were then cross-checked with two additional matrix-prediction algorithms, BIMAS (http://www-bimas.cit.nih.gov) and SYFPEITHI (http://www.syfpeithi.de). Peptides predicted by all three bioinformatic tools were selected and screened for similarity to the human genome using the NIH Blast server. Peptides showing homology to the human proteome were discarded.

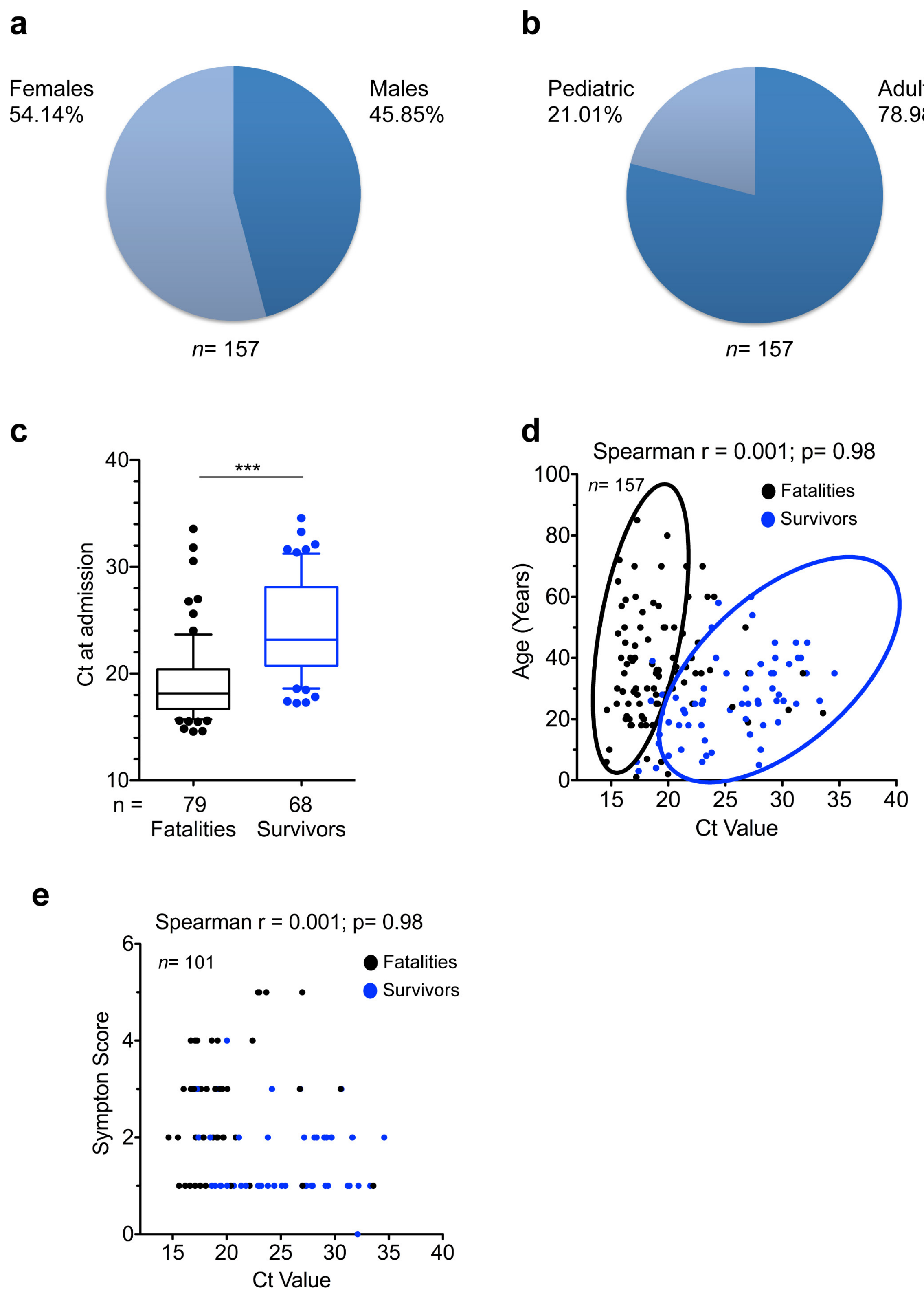
**Multiplex ELISA.** Concentrations of selected cytokines and chemokines were measured in 1:2 diluted plasma samples using Milliplex Map Human Cytokine/Chemokine Magnetic Bead Panel (EMD Millipore, Missouri, USA) on the Luminex platform LX200 (Luminex, Austin, USA). The procedure was performed according to the manufacturer's instructions.

**HLA genotyping.** High-resolution genotyping for HLA class I loci was performed by PCR-sequence-based typing, as recommended by the 13th International Histocompatibility Workshop (available at: http://www.ihwg.org). HLA sequences were analysed using the ASSIGN software (Conexio Genomics).

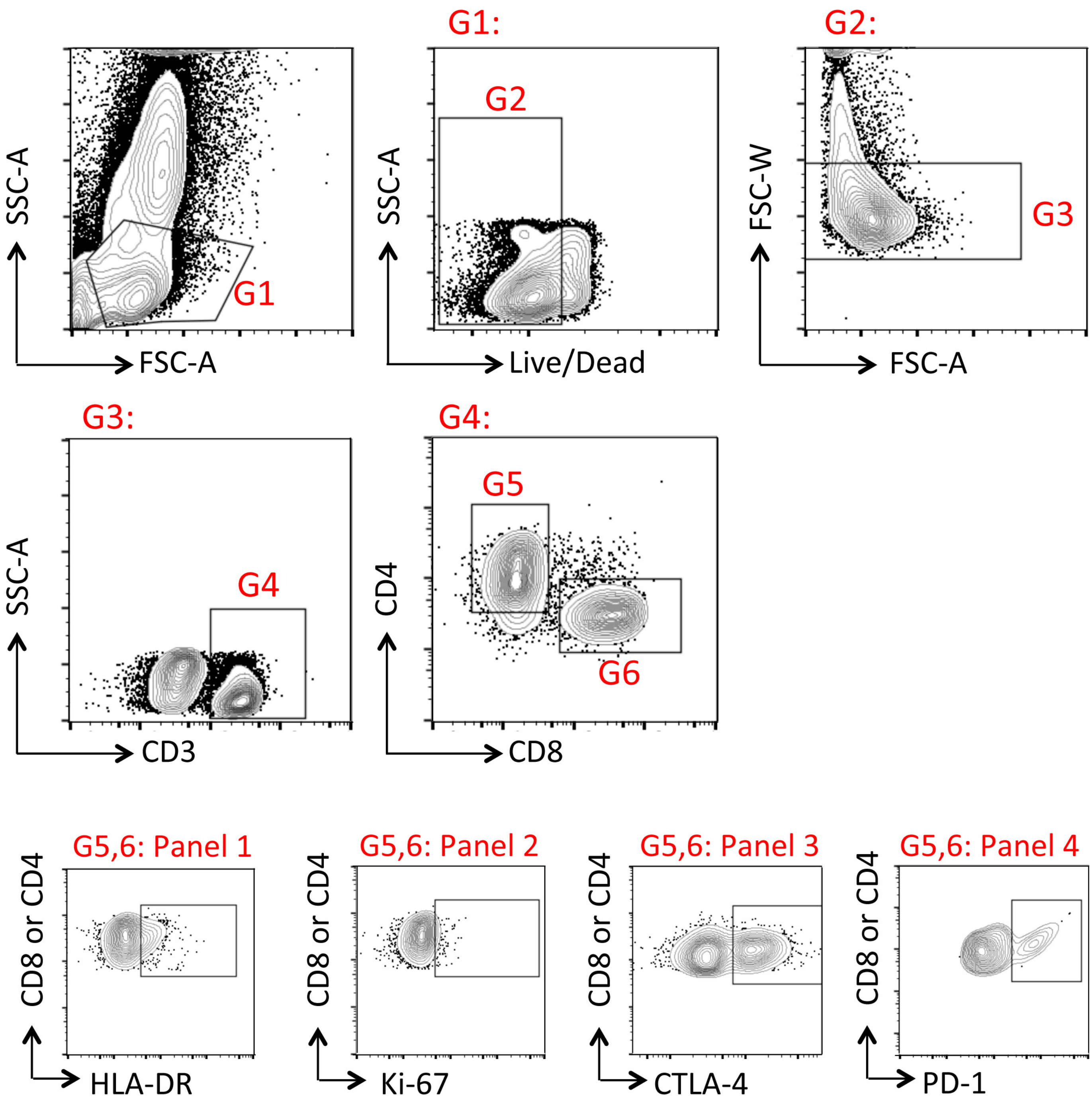
**Statistical analysis.** Non-parametric statistics and plots were performed in Graphpad Prism software as described in the figure legends. Sample distribution is illustrated throughout the manuscript using box-and-whisker plots. In all figures the boxes extend from the 25th to 75th percentiles and the horizontal bar is plotted as the median. The bars (whiskers) represent sample distribution down to the 10th percentile (lower bar) and up to the 90th percentile (upper bar).



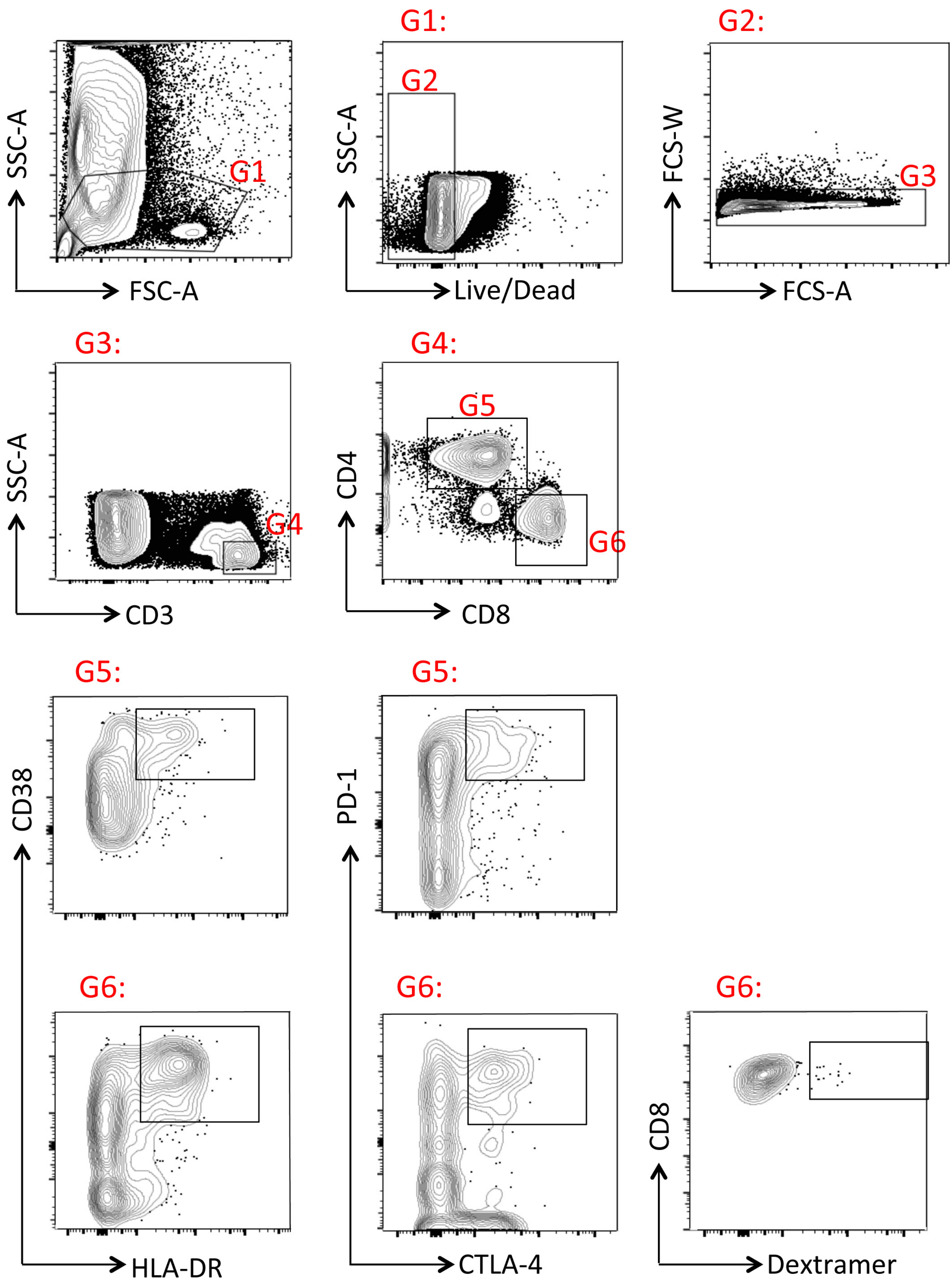

**Extended Data Figure 1 | Initial immunophenotyping data from Guéckédou. a**, Graph depicting the number of samples tested by the EMLab unit in Guéckédou by function of time since the beginning of the outbreak. The blue square indicates the period in which leftover whole blood samples from the diagnostic activities were shipped to the BSL-4 laboratory in Hamburg for initial immunophenotyping. **b**, Demographic data of the Guéckédou EVD patient cohort. Adults were $\geq 18$ years of age and paediatric patients were $<18$ years of age. **c**, Comparison of the expression of CTLA-4 assessed by median fluorescence intensity ratio (MFIR) in $CD4^+$ and $CD8^+$ T cells of EVD patients (POS, black boxes) and non-EVD controls (NEG, green boxes). MFIR represents the ratio between the CTLA-4-specific signal divided by the fluorescence minus one (FMO) signal of the same cell population. **d**, Comparison between CTLA-4 MFIR values in $CD8^+$ T cells from fatal (black) versus surviving (blue) EVD cases. In all panels, the ends of the whiskers in the box-and-whisker plots represent the 10th and 90th percentile, respectively. Statistical analysis was performed by non-parametric Mann–Whitney test; $*P \leq 0.05$; $***P \leq 0.0001$.



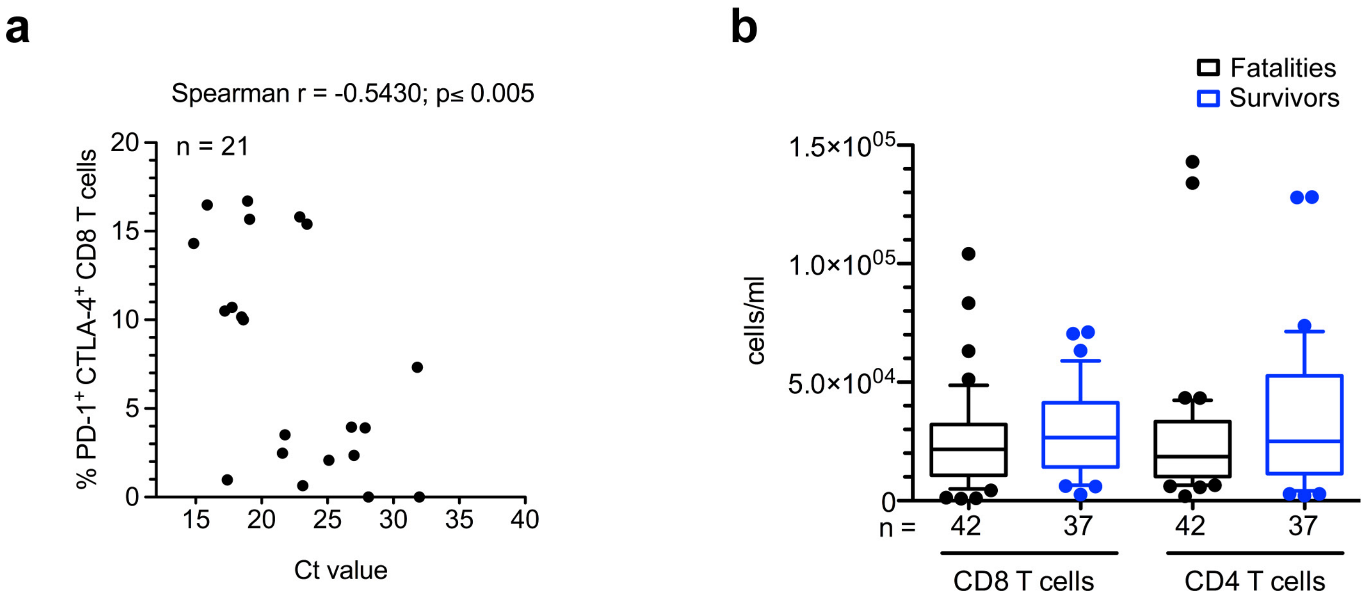

**Extended Data Figure 2 | Epidemiological data of patients tested by EMLab unit in Coyah. a**, **b**, Demographic data of the Coyah EVD patient cohort. Adults were ≤18 years of age and paediatric patients were <18 years of age. The median age of the 157 patients in the study was 26 years (interquartile range (IQR) 20–38 years). Percentages of males and females were comparable within all groups, with adults accounting for 79% of patients. **c**, Box-and-whisker plots depicting statistical association between $C_t$ values and outcome. The case–fatality ratio (CFR) was 51.6%. Fatalities and survivors were compared via non-parametric Mann–Whitney test; $^{***}P < 0.001$. **d**, Correlation between $C_t$ value and age of the patients. The $C_t$ value did not correlate with age. However, survivors clustered in a group characterized by $C_t$ value higher than 18 and by age less than 40 years (cluster encircled in blue). Statistical significance was tested by non-parametric Spearman correlation analysis. **e**, $C_t$ values correlated negatively with symptom scoring, so that low $C_t$ values were associated with severe disease symptoms. Symptom score was calculated as the summation of individual symptoms (bleeding, liver dysfunction, respiratory distress, kidney failure, neurological symptoms and anorexia) from '0' (no symptoms) to '6' (all symptoms present). In the box-and-whisker plot the ends of the whiskers represent the 10th and 90th percentile, respectively. Statistical analysis was performed by non-parametric Mann–Whitney test; $^{***}P \leq 0.0001$.




**Extended Data Figure 3 | Gating strategy for flow cytometry studies in Guinea.** All samples evaluated in the field were aliquoted for four panels. All panels had the following common gating: G1, lymphocyte gate; G2, live cells; G3, singlets; G4, T cells; G5, $CD4^+$ T cells; G6, $CD8^+$ T cells. Panels 1, 2, 3 and 4 evaluated expression of HLA-DR, Ki-67, CTLA-4 and PD-1, respectively, in either $CD4^+$ or $CD8^+$ T cells.




**Extended Data Figure 4 | Gating strategy for flow cytometry in Hamburg.** Cryopreserved PBMC samples from Coyah were thawed as indicated in the Methods. The following gates were used for sample analysis: G1, lymphocyte gate; G2, live cells; G3, singlets; G4, T cells; G5, $CD4^+$ T cells; G6, $CD8^+$ T cells. In G5 or G6, samples were evaluated for co-expression of the indicated cell markers. Dextramer staining was evaluated in G6 following protocols described in the Methods.



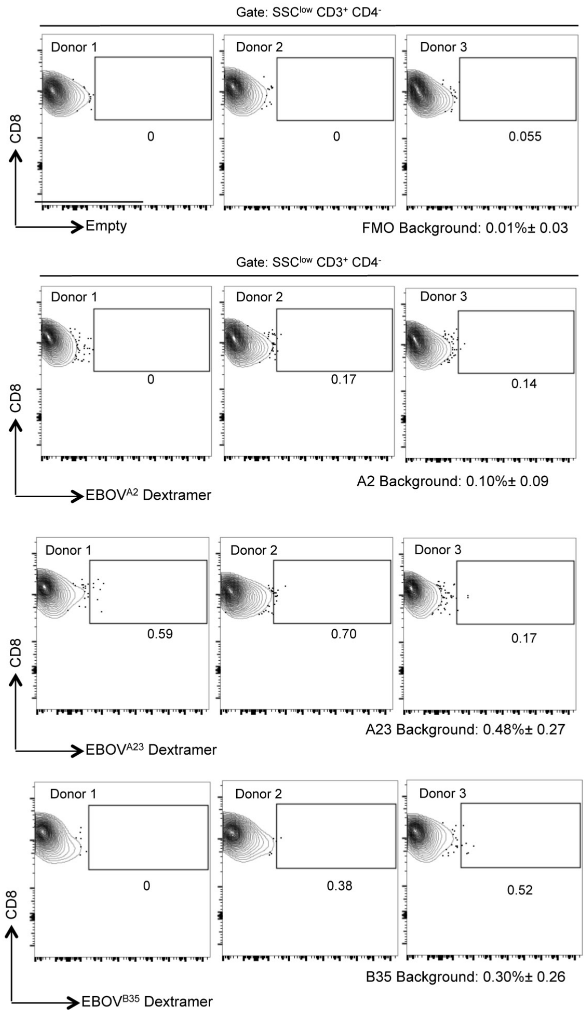

**Extended Data Figure 5 | Correlation of double positive PD-1$^+$/CTLA-4$^+$ CD8$^+$ T cells with $C_t$ values and lymphopenia. a**, Graph showing correlation between the frequency of CD8$^+$ T cells co-expressing PD-1 and CTLA-4 and the $C_t$ value. Correlation analysis was done via non-parametric Spearman correlation test. **b**, Box-and-whisker plots depicting the concentration of CD4$^+$ and CD8$^+$ T cells in blood of fatal and surviving EVD patients. The ends of the whiskers in the box-and-whisker plots represent the 10th and 90th percentile, respectively.

**a**

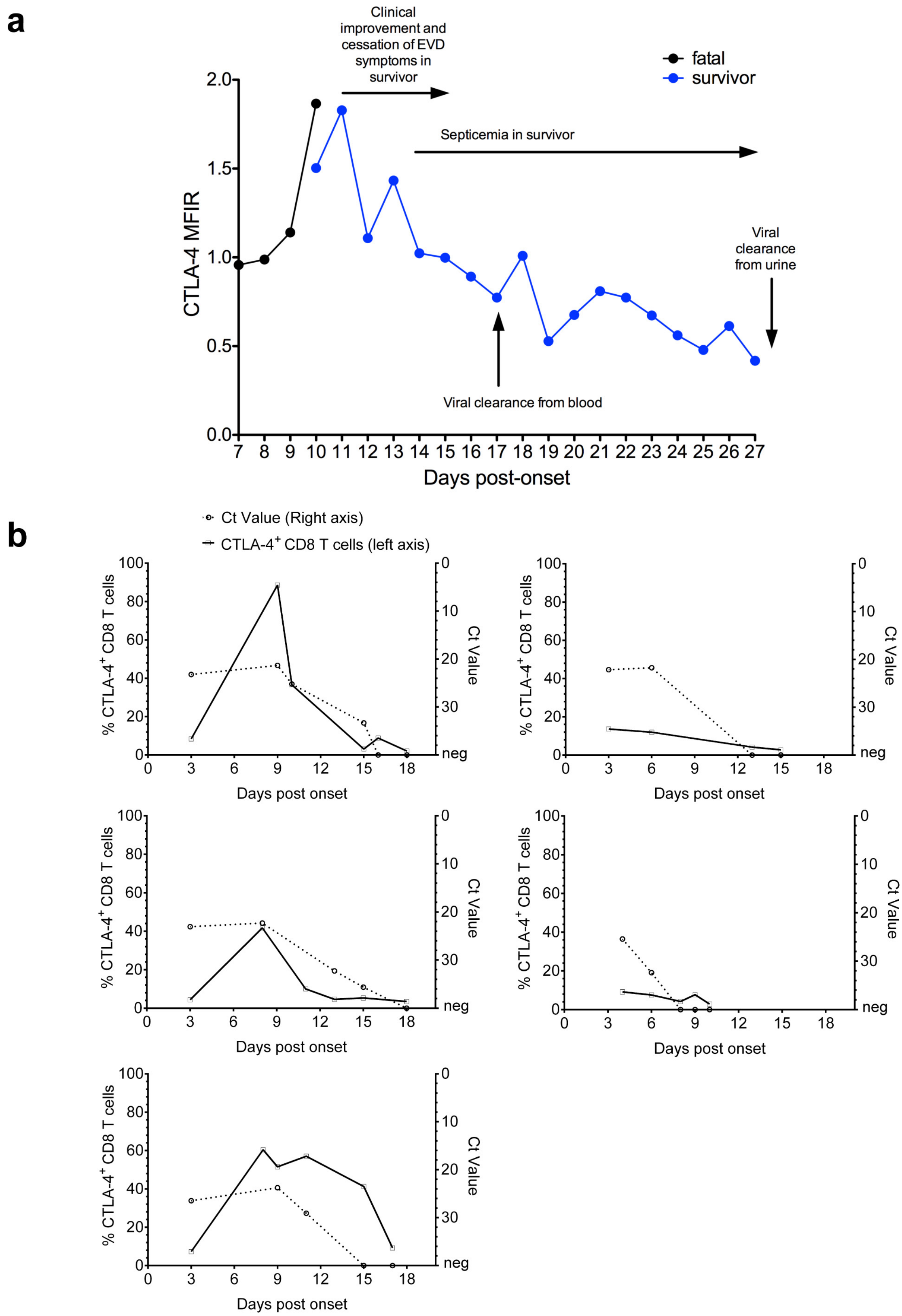
| | Aminoacid position in NP | | Sequence | $IC_{50}$ |
|---|---|---|---|---|
| HLA-A*02.01 | 74 | 82 | LLMLCLHHA | 15.6 |
| | 83 | 91 | YQGDYKLFL | 26.68 |
| | 116 | 124 | RLEELLPAV | 15.31 |
| | 150 | 158 | FLSFASLFL | 11.48 |
| | 202 | 210 | RLMRTNFLI | 11.88 |
| | 311 | 319 | GLFPQLSAI | 35.09 |
| | 404 | 412 | KLTEAITAA | 23.86 |
| | 669 | 677 | HMMKDEPVV | 40.46 |
| HLA-A*23.01 | 82 | 90 | AYQGDYKLF | 40 |
| HLA-B*35:01 | 313 | 321 | FPQLSAIAL | 7 |
| | 688 | 696 | YPDSLEEEY | 7 |
| | 179 | 187 | HAEQGLIQY | 12 |
| | 660 | 668 | GPFDAVLYY | 19 |
| | 58 | 66 | QAFEAGVDF | 21 |
| | 603 | 611 | TVAPPAPVY | 23 |
| | 421 | 429 | YDDDDDIPF | 27 |

Red: Previously published peptides with affinity for the indicated HLAs. $IC_{50}$= Half maximal inhibitory concentration (nM). Only peptides with predicted $IC_{50}$< 50nM were selected. Blue squares= Peptides chosen for dextramer design.

**b**




**Extended Data Figure 6 | *In silico* peptide analysis and dextramer design. a**, Selection of peptides consisting of nine amino acid residues corresponding to the EBOV nucleoprotein sequence predicted to bind the indicated HLA alleles. $IC_{50}$ values for peptide binding to HLA were predicted by the artificial neural network (ANN) at the Immune Epitope Database and Analysis Resource (IEDB) (http://www.iedb.org). **b**, Dextramer background was determined by staining of HLA-matched healthy donor peripheral blood leukocyte samples. T cells were gated as indicated in Extended Data Fig. 4. Plots in the upper row represent staining of a FMO (fluorescent minus one) sample in which the APC channel was left empty. Lower rows show background dextramer staining as indicated. The mean background staining plus minus standard deviation is indicated for each dextramer and the FMO.




**Extended Data Figure 7 | Longitudinal analysis of CTLA-4 expression in the CD8$^+$ T cell compartment during the course of EVD in two patients. a**, Graph depicts the levels of expression of CTLA-4 in CD8$^+$ T cells of a fatal versus a surviving EVD case over the course of the disease. Both patients were treated in Europe. Samples were taken at consecutive days starting immediately after patient admission as indicated. MFIR represents the ratio between the CTLA-4-specific signal divided by the fluorescence minus one (FMO) signal of the same cell population. **b**, Longitudinal analysis of CTLA-4 expression in CD8$^+$ T cells and $C_t$ values in survivors from Coyah.

doi:10.1038/nature17677

# A single injection of anti-HIV-1 antibodies protects against repeated SHIV challenges

Rajeev Gautam[1]*, Yoshiaki Nishimura[1]*, Amarendra Pegu[2], Martha C. Nason[3], Florian Klein[4,5,6], Anna Gazumyan[4], Jovana Golijanin[4], Alicia Buckler-White[1], Reza Sadjadpour[1], Keyun Wang[2], Zachary Mankoff[2], Stephen D. Schmidt[2], Jeffrey D. Lifson[7], John R. Mascola[2], Michel C. Nussenzweig[4,8] & Malcolm A. Martin[1]

**Despite the success of potent anti-retroviral drugs in controlling human immunodeficiency virus type 1 (HIV-1) infection, little progress has been made in generating an effective HIV-1 vaccine. Although passive transfer of anti-HIV-1 broadly neutralizing antibodies can protect mice or macaques against a single high-dose challenge with HIV or simian/human (SIV/HIV) chimaeric viruses (SHIVs) respectively[1–8], the long-term efficacy of a passive antibody transfer approach for HIV-1 has not been examined. Here we show, on the basis of the relatively long-term protection conferred by hepatitis A immune globulin, the efficacy of a single injection (20 mg kg$^{-1}$) of four anti-HIV-1-neutralizing monoclonal antibodies (VRC01, VRC01-LS, 3BNC117, and 10-1074 (refs 9–12)) in blocking repeated weekly low-dose virus challenges of the clade B SHIV$_{AD8}$. Compared with control animals, which required two to six challenges (median = 3) for infection, a single broadly neutralizing antibody infusion prevented virus acquisition for up to 23 weekly challenges. This effect depended on antibody potency and half-life. The highest levels of plasma-neutralizing activity and, correspondingly, the longest protection were found in monkeys administered the more potent antibodies 3BNC117 and 10-1074 (median = 13 and 12.5 weeks, respectively). VRC01, which showed lower plasma-neutralizing activity, protected for a shorter time (median = 8 weeks). The introduction of a mutation that extends antibody half-life into the crystallizable fragment (Fc) domain of VRC01 increased median protection from 8 to 14.5 weeks. If administered to populations at high risk of HIV-1 transmission, such an immunoprophylaxis regimen could have a major impact on virus transmission.**

It is now recognized that, unlike most other prophylactic vaccines for human viral pathogens, an effective vaccine against HIV-1 will probably need to completely block the establishment of a productive infection within a very short time frame (1–3 days of transmission). Such protection has, in fact, been achieved by administering polyclonal and monoclonal anti-HIV-1-neutralizing antibodies (NAbs) to humanized mice or macaques before challenge with SIV/HIV chimaeric viruses (SHIVs)[1–8].

During the past 7 years, monoclonal antibodies (MAbs) have been isolated from selected HIV-1 infected individuals, who generate anti-viral NAbs (bNAbs) with broad and potent activity against isolates of diverse genetic and geographical origin[13]. Several of these bNAbs have been used to suppress ongoing viral infections in humanized mice, macaques, and humans[14–18]. Pre-exposure immunoprophylaxis with bNAbs has also been evaluated in macaque models. In most of these experiments, a single dose of antibody, typically infused 24–48 h before a single high-dose virus challenge, was sufficient to block infection by a virus challenge, capable of establishing an infection in all untreated animals[4,19–21]. Humans, however, are usually exposed to much lower doses of virus on several occasions before becoming infected with HIV-1 (ref. 22).

It is worth noting that before the development of an effective hepatitis A virus vaccine, pre-exposure immunoprophylaxis with hepatitis A immune globulin was common practice for travellers to endemic regions of the world; protective effects lasted 3–5 months[23]. Prophylactic administration of antibodies against other microbial pathogens has also been used to prevent disease[24]. On the basis of this idea, we explored the possibility that a single administration of a potent neutralizing anti-HIV MAb, in the setting of repeated low-dose SHIV challenges, might protect for extended periods of time, thereby providing a proof of concept for periodic administration of MAb as an alternative to HIV-1 vaccination.

We initially selected three MAbs for the repeated low-dose SHIV challenge experiment on the basis of their previously described activity in blocking virus acquisition in a cohort of 60 macaques after a single high-dose SHIV challenge[21]. Two of these antibodies (VRC01 (ref. 12) and 3BNC117 (ref. 11)) target the gp120 CD4bs and one (10-1074 (ref. 10)) is dependent on the presence of HIV-1 gp120 N332 glycan, located immediately downstream of the V3 loop. The challenge virus selected for the present study was SHIV$_{AD8-EO}$ (ref. 25), an R5-tropic molecular-cloned derivative of the clade B SHIV$_{AD8}$ (ref. 26), which possesses multiple properties typical of pathogenic HIV-1 isolates[27].

When tested against large HIV-1 pseudovirus panels including multiple clades, 3BNC117 and VRC01 neutralize more than 80% of the viral isolates and 10-1074 neutralizes between 60% and 70%. Against sensitive viruses, 10-1074 is the most potent, followed by 3BNC117 and VRC01 (ref. 28). Consistent with this trend, the 50% inhibitory concentration (IC$_{50}$) values for VRC01, 3BNC117, and 10-1074 against SHIV$_{AD8-EO}$ were 0.67, 0.06, and 0.08 μg ml$^{-1}$, respectively, and the 80% (IC$_{80}$) values were 2.04, 0.19, and 0.18 μg ml$^{-1}$, respectively (Extended Data Fig. 1a). Neutralization sensitivities were also measured using the SHIV challenge stock in a single round of infection assay in TZM-bl cells, using replication-competent SHIV$_{AD8-EO}$. The IC$_{50}$ and IC$_{80}$ values for VRC01, 3BNC117, and 10-1074 in this assay system were 2.06, 0.12, and 0.05, and 7.14, 0.32, and 0.14 μg ml$^{-1}$, respectively (Extended Data Fig. 1b).

In an initial experiment designed to simulate low-dose mucosal transmission in humans, a cohort of nine monkeys was challenged weekly by the intrarectal route with ten 50% tissue culture infectious doses (TCID$_{50}$) of SHIV$_{AD8-EO}$, in the absence of antibody treatment.

[1]Laboratory of Molecular Microbiology, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, Maryland 20892, USA. [2]Vaccine Research Center, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, Maryland 20892, USA. [3]Biostatistics Research Branch, Division of Clinical Research, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, Maryland 20892, USA. [4]Laboratory of Molecular Immunology, The Rockefeller University, New York, New York 10065, USA. [5]Laboratory of Experimental Immunology, Center for Molecular Medicine Cologne (CMMC), University of Cologne, 50931 Cologne, Germany. [6]Department I of Internal Medicine, Center of Integrated Oncology Cologne-Bonn, University Hospital Cologne, 50937 Cologne, Germany. [7]AIDS and Cancer Virus Program, Leidos Biomedical Research, Frederick National Laboratory for Cancer Research, Frederick, Maryland 21702, USA. [8]Howard Hughes Medical Institute, Chevy Chase, Maryland 20815, USA.
*These authors contributed equally to this work.

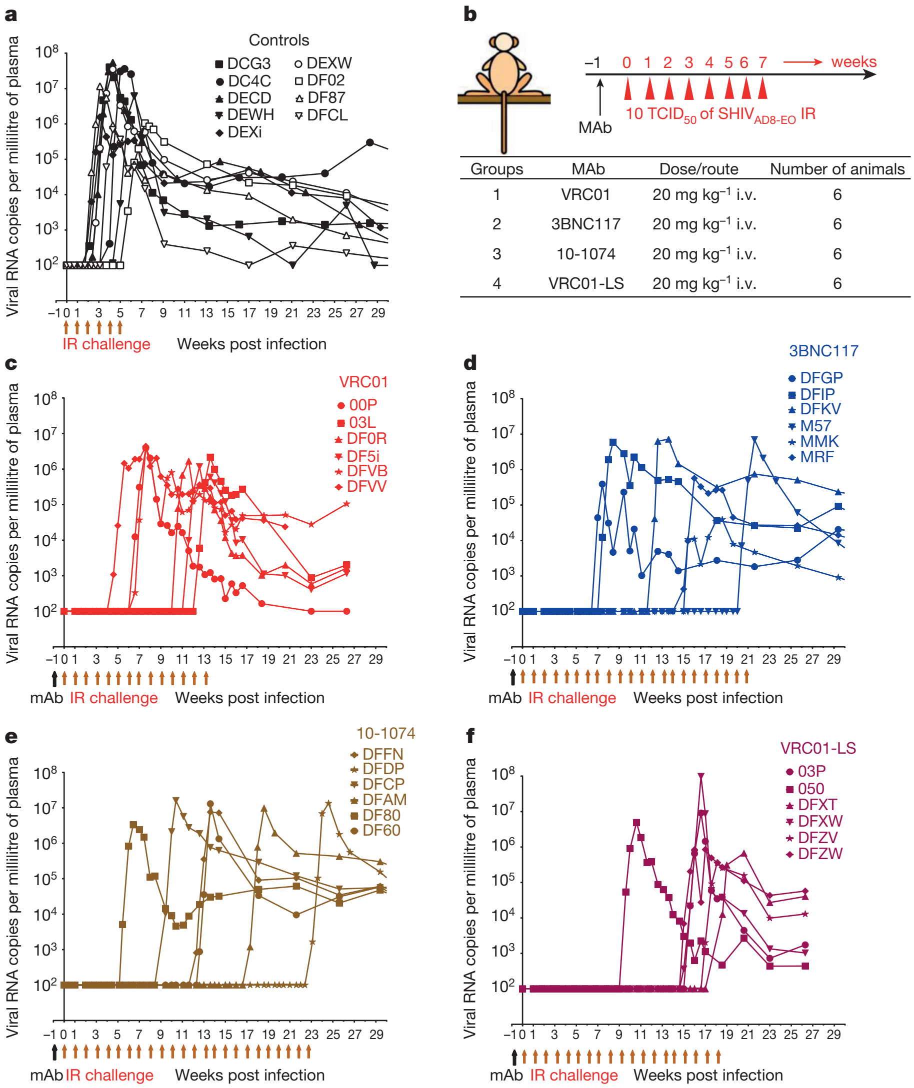



**Figure 1 | HIV MAbs delay virus acquisition after repeated low-dose intrarectal SHIV$_{AD8-EO}$ challenges. a**, Plasma viral loads in macaques receiving no MAb (controls, $n=9$). IR, intrarectal. **b**, Representation of the regimen used to assess the protective efficacy of MAbs. Macaques were intravenously (i.v.) administered the indicated MAbs at a dose of 20 mg kg$^{-1}$and challenged 1 week later and every week thereafter. **c–f**, Plasma viral loads in macaques administered VRC01, 3BNC117, 10-1074, and VRC01-LS bNAbs, respectively.

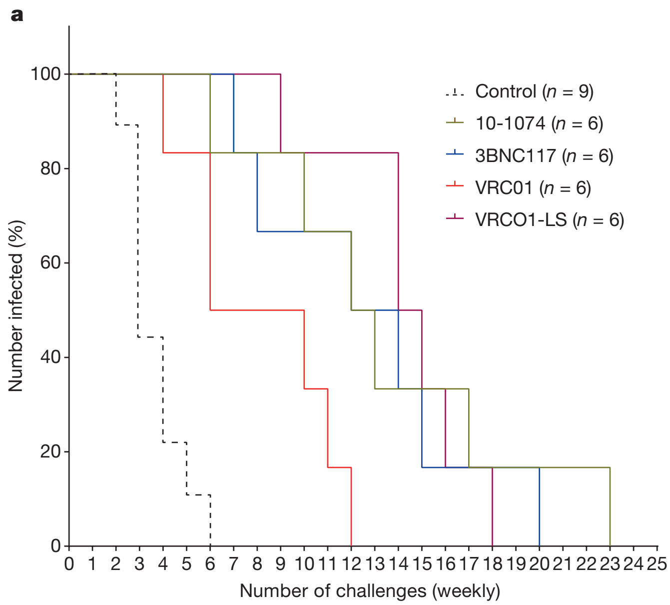
As shown in Fig. 1a, plasma viraemia became detectable after two to six challenges, with a median of 3.0 weekly virus exposures needed to infect all nine animals. On the basis of these results, the inoculum size administered to each monkey per challenge was estimated to be 0.27 50% animal infectious doses (AID$_{50}$).

The regimen used to assess the protective efficacy of the three anti-HIV-1 MAbs against a repeated low-dose rectal challenge of SHIV$_{AD8-EO}$ is shown in Fig. 1b. Individual MAbs (20 mg kg$^{-1}$) were administered a single time intravenously to three cohorts of six animals. Starting 1 week later, each group was challenged weekly by the intrarectal route with ten TCID$_{50}$ of SHIV$_{AD8-EO}$. Samples of blood, collected at regular intervals, were monitored for levels of viral RNA, concentrations of MAb, and anti-SHIV-neutralizing titres. The number of virus challenges required to establish a SHIV$_{AD8-EO}$ infection, indicated by measurable viraemia (>100 viral RNA copies per millilitre of plasma), in the recipients of the anti-HIV-1 MAbs was compared with that needed for virus acquisition in the control group.

In all cases, the administration of MAbs delayed virus acquisition. Animals receiving VRC01 required 4–12 challenges; 3BNC117 required 7–20 challenges; and 10-1074, 6–23 challenges (Fig. 1c–e). The differences in the number of challenges required for infection, and thus the median times to virus acquisition compared with control monkeys, were 8 weeks for VRC01, 13 weeks for 3BNC117, and 12.5 weeks for 10-1074.

The pharmacokinetic profile of VRC01 was altered by introducing two amino-acid mutations (M428L and N434S, referred to as 'LS') into its Fc domain, which increased its half-life in both plasma and tissues[9]. The neutralization activity of this VRC01 derivative, designated VRC01-LS, was first tested in the TZM-bl assay and, as expected, it exhibited IC$_{50}$ and IC$_{80}$ values similar to VRC01 (Extended Data Fig. 1a, b). When administered to six macaques in the previously described repeated low-dose SHIV$_{AD8-EO}$ challenge system, the VRC01-LS-treated animals required 9–18 challenges (median = 14.5) for all of the monkeys to become infected (Fig. 1f). Thus the modified VRC01-LS Fc domain conferred an estimated 1.8-fold increase in the number of challenges, resulting in successful acquisition compared with the parental VRC01 MAb.

The protective effects of the four anti-HIV-1 MAbs are described by Kaplan–Meier analysis in which the percentage of macaques remaining uninfected is plotted against the number of SHIV$_{AD8EO}$ challenges (Fig. 2a). Significantly increased numbers of challenges were required to establish infections in the recipients of the VRC01, 3BNC117, 10-1074, and VRC01-LS MAbs than in the control animals ($P=0.007$, 0.002, 0.002, and 0.002, respectively), using the Wilcoxon rank-sum test




**b**

| Groups compared | | Wilcoxon *P* value |
|---|---|---|
| Control | VRC01 | 0.007 |
| Control | 3BNC117 | 0.002 |
| Control | 10-1074 | 0.002 |
| Control | VRC01-LS | 0.002 |
| VRC01 | 3BNC117 | 0.077 |
| VRC01 | 10-1074 | 0.106 |
| VRC01 | VRC01-LS | 0.02 |
| 10-1074 | 3BNC117 | 0.5 |

**Figure 2 | Kaplan–Meier analysis and magnitude of protection by HIV MAbs in repeated low-dose challenge. a**, Kaplan–Meier survival curves for recipients of the four bNAbs and the cohort of control animals. The percentage of macaques remaining uninfected is plotted against the number of ten $TCID_{50}$ $SHIV_{AD8\text{-}EO}$ intrarectal challenges required to establish infections. **b**, Statistical differences are represented as *P* values (Wilcoxon rank-sum test) by comparing the number of challenges resulting in infection between control animals and an individual MAb recipient group or between different MAb-treated groups.

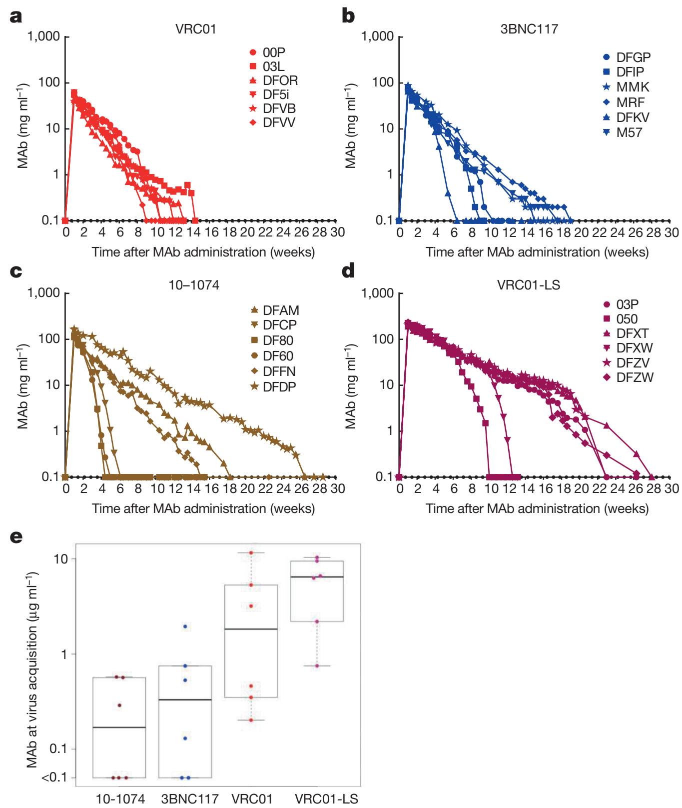



**Figure 3 | Plasma concentrations of the infused MAbs in macaques correlate with long-term protection from SHIV infection. a–d**, Plasma antibody concentrations in macaques administered VRC01, 3BNC117, 10-1074, and VRC01-LS decay over time. **e**, Median plasma concentrations at the times of virus breakthrough in bNAb recipients were 0.169 (10-1074), 0.330 (3BNC117), 1.825 (VRC01), and 6.446 (VRC01-LS), respectively. Boxes represent the twenty-fifth and seventy-fifth percentiles, and the heavier line represents the median value for each group. The top and bottom horizontal bars outside the boxes represent the maximum and minimum of the data, respectively.

(Fig. 2b). A comparison of the individual pairs of Kaplan–Meier curves revealed that 10-1074, 3BNC117, and VRC01-LS were not significantly different from each other in blocking infection.

Ultrasensitive nested quantitative reverse-transcription PCR (qRT–PCR) and qPCR assays for plasma viral RNA and cell-associated viral RNA and DNA[29] were performed on plasma and peripheral blood mononuclear cell samples, collected from recipients of the different neutralizing MAbs, before $SHIV_{AD8\text{-}EO}$ breakthrough infections, as assessed by plasma viraemia measured in our standard assay. In all cases, the levels of viral RNA and DNA measured with the ultrasensitive assays were below detectable limits (Extended Data Table 1).

The plasma concentrations of the infused MAbs were measured longitudinally in individual animals beginning 1 week after infusion. (Fig. 3 and Extended Data Tables 2 and 3). The median plasma concentrations at the times of virus breakthrough for the 10-1074 and 3BNC117 recipient cohorts were 0.169 and 0.330 $\mu g\, ml^{-1}$, respectively (Fig. 3e). These values are comparable to the $IC_{80}$ values determined *in vitro*, using the TZM-bl assay with replication competent $SHIV_{AD8\text{-}EO}$ (Extended Data Fig. 1b). The median plasma concentrations at the times of virus acquisition for VRC01 and VRC01-LS were 10- to 20-fold higher (1.825 and 6.446 $\mu g\, ml^{-1}$) and were also in the same range as the $IC_{80}$ values determined *in vitro* (Extended Data Fig. 1b). It is worth noting that three of the six recipients of the 10-1074 MAb experienced rapid decay of plasma antibody, which fell to background levels between weeks 4 and 6 after administration (Fig. 3c). A similar pattern occurred for three of the 3BNC117 MAb and VRC01-LS recipients, although the decline of antibody in plasma was delayed in these two groups animals (Fig. 3b). This rapid clearance of plasma MAbs in the subgroups of the 10-1074 and 3BNC117 recipients tracked with the emergence of anti-antibody responses to the infused anti-HIV-1 human MAbs (Extended Data Fig. 2). For monkeys infused with the 10-1074 MAb, the median number of challenges for successful infection in the three-animal subgroup not experiencing the rapid anti-antibody induced decay was 17.0 weeks compared with 12.5 for the entire 10-1074 recipient cohort.

Probit analysis was also used to estimate the probability of infection as a function of the imputed plasma MAb concentration at the time of each challenge. The probability of infection per infection for the control monkeys was 0.27, estimated by pooling all of the $SHIV_{AD8\text{-}EO}$ challenges to this group of animals; this is indicated by the single open circle along the ordinate of Extended Data Fig. 3. Not unexpectedly, the curves relating antibody concentration and virus acquisition for VRC01 and VRC01-LS were superimposed on one another even though VRC01-LS had a longer half-life *in vivo*. In this same analysis, the curves for 10-1074 and 3BNC117, which conferred lower probabilities for infection at each plasma MAb concentration, reflected their greater neutralization potency against the challenge virus, relative to the VRC01 antibodies. At a 10-1074 MAb plasma concentration of 1 $\mu g\, ml^{-1}$, the model predicts a probability of infection, for a single challenge, of 0.044, approximately sixfold less than that estimated for animals receiving no antibodies.

**a**

| | | Week 1 | Week 2 | Week 4 | Week 6 | Week 8 | Week 12 | Week 16 | Week 20 | Week 24 | Week 28 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| VRC01 | 00P | 422 | 155 | 95 | 63 | 49 | 21 | <20 | | | |
| | 03L | 430 | 148 | 63 | 42 | 30 | <20 | | | | |
| | DF0R | 303 | 68 | 30 | 29 | 21 | <20 | | | | |
| | DF5i | 474 | 95 | 54 | 39 | 26 | <20 | | | | |
| | DFVB | 401 | 100 | 52 | 40 | 33 | <20 | | | | |
| | DFVV | 367 | 190 | 129 | 75 | 25 | <20 | | | | |
| 3BNC117 | DFGP | 2720 | 499 | 379 | 114 | 48 | <20 | | | | |
| | DFIP | 3681 | 849 | 467 | 220 | 28 | <20 | | | | |
| | DFKV | 3526 | 510 | 295 | <20 | <20 | | | | | |
| | M57 | 2971 | 486 | 322 | 135 | 48 | <20 | | | | |
| | MMK | 4125 | 1380 | 576 | 289 | 41 | <20 | | | | |
| | MRF | 2244 | 1082 | 553 | 173 | 100 | <20 | | | | |
| 10-1074 | DF60 | 5145 | 1750 | 31 | <20 | | | | | | |
| | DF80 | 7266 | 1799 | <20 | | | | | | | |
| | DFAM | 5540 | 2674 | 1033 | 335 | 158 | 39 | <20 | | | |
| | DFCP | 9057 | 2688 | 248 | <20 | | | | | | |
| | DFDP | 9168 | 3853 | 1399 | 811 | 443 | 149 | 59 | 31 | 25 | <20 |
| | DFFN | 7041 | 2346 | 532 | 177 | 61 | 22 | <20 | | | |
| VRC01-LS | 03P | 1218 | 893 | 856 | 316 | 232 | 87 | <20 | | | |
| | 50 | 1082 | 694 | 418 | 218 | 51 | 25 | <20 | | | |
| | DFXT | 1680 | 480 | 398 | 378 | 229 | 100 | 55 | <20 | | |
| | DFXW | 2334 | 682 | 431 | 430 | 192 | <20 | <20 | | | |
| | DFZV | 1504 | 560 | 487 | 318 | 220 | 136 | 98 | <20 | | |
| | DFZW | 1473 | 674 | 590 | 289 | 219 | 123 | 78 | <20 | | |

**b**

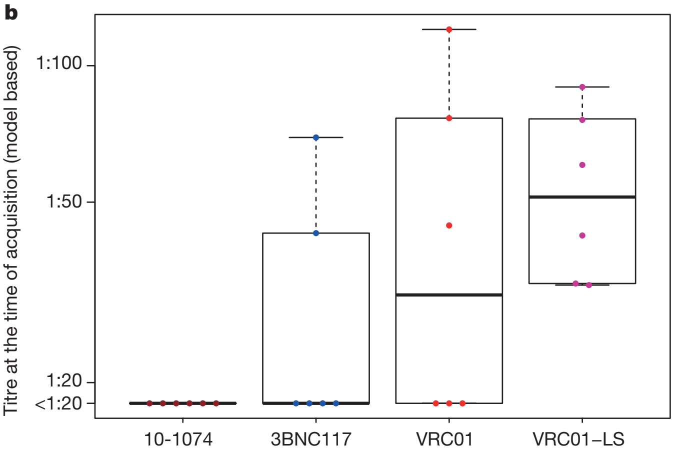



**Figure 4 | The decline of neutralizing antibody titres in plasma over time in macaques corresponds to the time of virus acquisition. a**, Plasma $IC_{50}$ titres of the indicated MAbs were determined longitudinally using the TZM-bl cell assay. $IC_{50}$ values are colour coded: 21–99 as green; 100–999 as yellow; and ≥1,000 as red. **b**, Plasma-neutralizing titres at the time of virus acquisition for the four groups of MAb recipients. Boxes represent the twenty-fifth and seventy-fifth percentiles, and the heavier line represents the median value for each group. The top and bottom horizontal bars outside of the boxes represent the maximum and minimum of the data, respectively.

The plasma neutralization titre was also determined for each of the MAb recipients at multiple times after infusion (Fig. 4a). The median plasma-neutralizing titres for the four groups of macaques at the time of $SHIV_{AD8-EO}$ acquisition were low: <1:20 (below the level of detection) for 10-1074 and 3BNC117 recipients; 1:27 for the VRC01 group; and 1:51 for the VRC01-LS cohort (Fig. 4b). As noted earlier for plasma MAb concentrations, the levels of detectable neutralizing activity in members of each cohort inversely correlated with the emergence of anti-antibodies (compare Fig. 4a and Extended Data Fig. 2).

In conclusion a single administration of potent anti-HIV-1-neutralizing MAbs to naive macaques was protective against repeated low-dose SHIV infection for several months. The duration of protection was directly related to antibody potency and half-life. When considered in the context of a potential exposure to HIV-1 in regions of the world where the HIV-1 is endemic, the barrier to infection when antibody concentrations remain above protective levels in infused individuals could have a profound impact on virus transmission. As noted earlier, anti-antibodies directed against some of the administered MAbs emerged quite rapidly in some macaques, and diminished their prophylactic efficacy. However, this is not likely to occur in humans as reported in a recent study of VRC01 (ref. 30). On the basis of the results obtained with VRC01 and VRC01-LS, it is also anticipated that the creation and use of 3BN117 and/or 10-1074 derivatives with the LS mutation should exhibit increased durability *in vivo*, resulting in protection of up to 6 months against $SHIV_{AD8-EO}$-infected macaques. The administration of a multivalent cocktail of these anti-viral bNAbs could augment their efficacy by increasing overall breadth and their capacity to block the transmission of resistant HIV-1 strains.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 10 February; accepted 17 March 2016.**

**Published online 27 April 2016.**

1. Baba, T. W. *et al.* Human neutralizing monoclonal antibodies of the IgG1 subtype protect against mucosal simian-human immunodeficiency virus infection. *Nature Med.* **6,** 200–206 (2000).
2. Hessell, A. J. *et al.* Effective, low-titer antibody protection against low-dose repeated mucosal SHIV challenge in macaques. *Nature Med.* **15,** 951–954 (2009).
3. Mascola, J. R. *et al.* Protection of macaques against vaginal transmission of a pathogenic HIV-1/SIV chimeric virus by passive infusion of neutralizing antibodies. *Nature Med.* **6,** 207–210 (2000).
4. Moldt, B. *et al.* Highly potent HIV-specific antibody neutralization *in vitro* translates into effective protection against mucosal SHIV challenge *in vivo*. *Proc. Natl Acad. Sci. USA* **109,** 18921–18925 (2012).
5. Nishimura, Y. *et al.* Determination of a statistically valid neutralization titer in plasma that confers protection against simian-human immunodeficiency virus challenge following passive transfer of high-titered neutralizing antibodies. *J. Virol.* **76,** 2123–2130 (2002).
6. Parren, P. W. *et al.* Antibody protects macaques against vaginal challenge with a pathogenic R5 simian/human immunodeficiency virus at serum levels giving complete neutralization *in vitro*. *J. Virol.* **75,** 8340–8347 (2001).
7. Pietzsch, J. *et al.* A mouse model for HIV-1 entry. *Proc. Natl Acad. Sci. USA* **109,** 15859–15864 (2012).
8. Shibata, R. *et al.* Neutralizing antibody directed against the HIV-1 envelope glycoprotein can completely block HIV-1/SIV chimeric virus infections of macaque monkeys. *Nature Med.* **5,** 204–210 (1999).
9. Ko, S. Y. *et al.* Enhanced neonatal Fc receptor function improves protection against primate SHIV infection. *Nature* **514,** 642–645 (2014).
10. Mouquet, H. *et al.* Complex-type N-glycan recognition by potent broadly neutralizing HIV antibodies. *Proc. Natl Acad. Sci. USA* **109,** E3268–E3277 (2012).
11. Scheid, J. F. *et al.* Sequence and structural convergence of broad and potent HIV antibodies that mimic CD4 binding. *Science* **333,** 1633–1637 (2011).
12. Zhou, T. *et al.* Structural basis for broad and potent neutralization of HIV-1 by antibody VRC01. *Science* **329,** 811–817 (2010).
13. Burton, D. R. & Mascola, J. R. Antibody responses to envelope glycoproteins in HIV-1 infection. *Nature Immunol.* **16,** 571–576 (2015).
14. Barouch, D. H. *et al.* Therapeutic efficacy of potent neutralizing HIV-1-specific monoclonal antibodies in SHIV-infected rhesus monkeys. *Nature* **503,** 224–228 (2013).
15. Caskey, M. *et al.* Viraemia suppressed in HIV-1-infected humans by broadly neutralizing antibody 3BNC117. *Nature* **522,** 487–491 (2015).
16. Klein, F. *et al.* HIV therapy by a combination of broadly neutralizing antibodies in humanized mice. *Nature* **492,** 118–122 (2012).
17. Shingai, M. *et al.* Antibody-mediated immunotherapy of macaques chronically infected with SHIV suppresses viraemia. *Nature* **503,** 277–280 (2013).
18. Lynch, R. M. *et al.* Virologic effects of broadly neutralizing antibody VRC01 administration during chronic HIV-1 infection. *Sci. Transl. Med.* **7,** 319ra206 (2015).
19. Rudicell, R. S. *et al.* Enhanced potency of a broadly neutralizing HIV-1 antibody *in vitro* improves protection against lentiviral infection *in vivo*. *J. Virol.* **88,** 12669–12682 (2014).
20. Saunders, K. O. *et al.* Sustained delivery of a broadly neutralizing antibody in nonhuman primates confers long-term protection against simian/human immunodeficiency virus infection. *J. Virol.* **89,** 5895–5903 (2015).
21. Shingai, M. *et al.* Passive transfer of modest titers of potent and broadly neutralizing anti-HIV monoclonal antibodies block SHIV infection in macaques. *J. Exp. Med.* **211,** 2061–2074 (2014).
22. Patel, P. *et al.* Estimating per-act HIV transmission risk: a systematic review. *AIDS* **28,** 1509–1519 (2014).
23. Fiore, A. E., Wasley, A. & Bell, B. P. Prevention of hepatitis A through active or passive immunization: recommendations of the Advisory Committee on Immunization Practices (ACIP). *Morbid. Mortal. Wkly Rep.* **55,** 1–23 (2006).
24. Graham, B. S. & Ambrosino, D. M. History of passive antibody administration for prevention and treatment of infectious diseases. *Curr. Opin. HIV AIDS* **10,** 129–134 (2015).

25. Shingai, M. *et al.* Most rhesus macaques infected with the CCR5-tropic SHIV(AD8) generate cross-reactive antibodies that neutralize multiple HIV-1 strains. *Proc. Natl Acad. Sci. USA* **109,** 19769–19774 (2012).
26. Nishimura, Y. *et al.* Generation of the pathogenic R5-tropic simian/human immunodeficiency virus SHIVAD8 by serial passaging in rhesus macaques. *J. Virol.* **84,** 4769–4781 (2010).
27. Gautam, R. *et al.* Pathogenicity and mucosal transmissibility of the R5-tropic simian/human immunodeficiency virus SHIV(AD8) in rhesus macaques: implications for use in vaccine studies. *J. Virol.* **86,** 8516–8526 (2012).
28. West, A. P. Jr *et al.* Computational analysis of anti-HIV-1 antibody neutralization panel data to identify potential functional epitope residues. *Proc. Natl Acad. Sci. USA* **110,** 10598–10603 (2013).
29. Hansen, S. G. *et al.* Profound early control of highly pathogenic SIV by an effector memory T-cell vaccine. *Nature* **473,** 523–527 (2011).
30. Ledgerwood, J. E. *et al.* Safety, pharmacokinetics and neutralization of the broadly neutralizing HIV-1 human monoclonal antibody VRC01 in healthy adults. *Clin. Exp. Immunol.* **182,** 289–301 (2015).

**Acknowledgements** We thank R. Plishka, A. Peach, and T. Lewis for determining plasma viral RNA loads, and K. Rice, R. Engel, R. Petros, and S. Fong for assisting in the maintenance of animals and assisting with procedures. We also thank R. Schwartz for clinical-grade VRC01 and VRC01-LS, and X. Chen for protein reagents for ELISA. We thank the National Institutes of Health (NIH) AIDS Research and Reference Reagent Program for TZM-bl cells. We thank R. Fast for ultrasensitive plasma SIV RNA assays and W. Bosche and M Hull for ultrasensitive peripheral blood mononuclear cell SIV RNA/DNA assays. This work was supported by the Intramural Research Program of the National Institute of Allergy and Infectious Diseases, NIH and, in part, with federal funds from the National Cancer Institute, NIH, under contract number HHSN261200800001E (to J.D.L.). The research was also funded in part by the Bill and Melinda Gates Foundation Collaboration for AIDS Vaccine Discovery Grants OPP1033115 and OPP1092074 (to M.C.Nu.), by the NIH under award numbers AI-100148, UM1 AI100663-01. M.C.Nu. is supported by the Robertson Foundation and the The Howard Hughes Medical Institute.

**Author Contributions** R.G., Y.N., M.A.M., M.C.Nu., and J.R.M. designed experiments; R.G., Y.N., A.P., F.K., A.G., J.G., A.B.W., R.S., K.W., Z.M., and S.D.S. performed experiments; R.G., Y.N., M.C.Na., M.A.M., M.C.Nu., J.R.M., and J.D.L. analysed data; R.G., Y.N., M.A.M., M.C.Nu., J.R.M., and J.D.L. wrote the manuscript.

**Author Information** Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.A.M. (malm@nih.gov).

## METHODS

**Animal experiments.** Thirty-three male and female rhesus macaques (*Macaca mulatta*) of Indian genetic origin from 2 to 4 years of age were housed and cared for in accordance with Guide for Care and Use of Laboratory Animals Report number NIH 82-53 (Department of Health and Human Services, Bethesda, Maryland, 1985) in a biosafety level 2 National Institute of Allergy and Infectious Diseases (NIAID) facility. All animal procedures and experiments were performed according to protocols approved by the Institutional Animal Care and Use Committee of NIAID, NIH. Animals were not randomized and the data collected were not blinded. Phlebotomies, euthanasia, and sample collection were performed as previously described[31]. All of the macaques used in this study were negative for the major histocompatibility complex (MHC) class I *Mamu-A*01*, *Mamu-B*08*, and *Mamu-B*17* alleles. No animals were excluded from the analysis.

**Antibodies.** The VRC01, 3BNC117, 10-1074, and VRC01-LS anti-HIV-1 monoclonal NAbs were isolated and produced as described elsewhere[9–12]. MAb 10-1074 was produced by transient transfection of IgH and IgL expression plasmids into the human embryonic kidney cells whereas VRC01, VRC01-LS, and 3BNC117 were produced from Chinese hamster ovary cells. All of the MAbs were IgG1. All of the monoclonal antibodies were purified by chromatography and sterile filtration and were endotoxin free. A single dose ($20\,\mathrm{mg\,kg^{-1}}$) of each MAb was administered intravenously to individual animals in four cohorts of monkeys.

**Virus challenge.** The origin and preparation of the tissue-culture-derived $\mathrm{SHIV_{AD8\text{-}EO}}$ stock have been previously described[25]. One week after MAb infusion, animals were challenged intrarectally with ten $\mathrm{TCID_{50}}$ of $\mathrm{SHIV_{AD8\text{-}EO}}$, and every week thereafter, until a virus infection was established. A paediatric nasal speculum was used to gently open the rectum and a 1 ml suspension of virus was slowly infused into rectal cavity using a plastic tuberculin syringe. An intrarectal challenge $\mathrm{SHIV_{AD8\text{-}EO}}$ inoculum size of ten $\mathrm{TCID_{50}}$ was chosen for repeated low-dose experiments on the basis of previous results indicating that (1) 1,000 $\mathrm{TCID_{50}}$ of $\mathrm{SHIV_{AD8\text{-}EO}}$ administered by the intrarectal route resulted in the establishment of infections of 30 of 30 rhesus monkeys and (2) an intrarectal virus titration suggested that 1,000 $\mathrm{TCID_{50}}$ of $\mathrm{SHIV_{AD8\text{-}EO}}$ was equivalent to approximately ten $\mathrm{AID_{50}}$ (ref. 21).

**Quantification of viral nucleic acids.** Viral RNA levels in plasma were determined by qRT–PCR (ABI Prism 7900HT sequence detection system; Applied Biosystems) as previously described[31]. Ultrasensitive measurement of plasma SIV gag RNA was performed as described, and cell-associated levels of SIV RNA and DNA were determined by a nested, hybrid real-time/digital PCR assay, essentially as reported previously[29].

**Antibody concentrations in plasma.** Plasma antibody levels were quantified by ELISA using purified MAbs as a standard and anti-antibody responses in plasma were also evaluated as reported earlier[9]. These assays were performed twice.

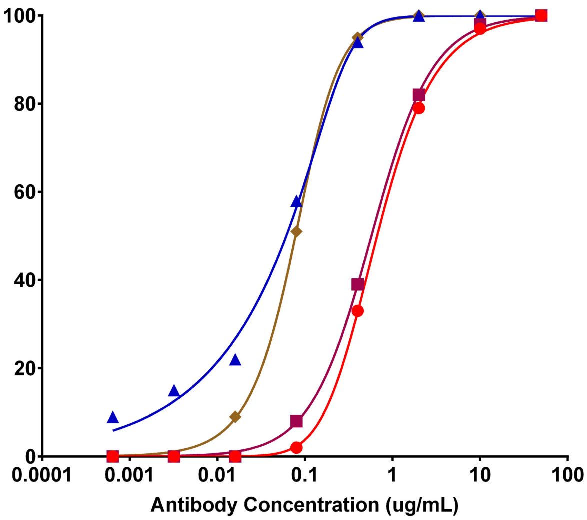
**Neutralization assays.** The titres of each MAb against $\mathrm{SHIV_{AD8\text{-}EO}}$ was assessed by two types of *in vitro* neutralization assay: (1) TZM-bl entry assay with pseudotype challenge virus[25,27] and (2) a single-round TZM-bl infectivity assay with replication competent challenge virus[32]. Antibody concentrations required to inhibit infection by 50% or 80% are reported as $\mathrm{IC_{50}}$ or $\mathrm{IC_{80}}$, respectively. TZM-bl cells were obtained through the NIH AIDS Reagent Program, Division of AIDS, NIAID, NIH, from J. C. Kappes, X. Wu and Tranzyme[33].These cells were not authenticated for this study and not tested for mycoplasma contamination. The neutralization activity present in plasma samples collected from rhesus macaques was assessed by TZM-bl entry assay with pseudotype challenge virus. The $\mathrm{IC_{50}}$ titre was calculated as the plasma dilution causing 50% reduction in RLUs compared with virus controls. The neutralization assays were repeated twice.

**Statistical analyses.** No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

A Wilcoxon rank-sum test was used to compare number of challenges until infection between each MAb group and control; these comparisons were considered primary and were compared with a Bonferroni-adjusted $\alpha$ of $0.05/4 = 0.0125$ to determine significance. Comparisons between antibodies were considered secondary and not adjusted for multiple comparisons. Finally, probit models were used to model the probability of infection at each challenge as a function of concurrent antibody concentration. Since these values were not always measured at the precise time of challenge, antibody concentrations were modelled separately for each animal over time, and these models were used to impute the concentration at the exact time of each challenge for the probit model.

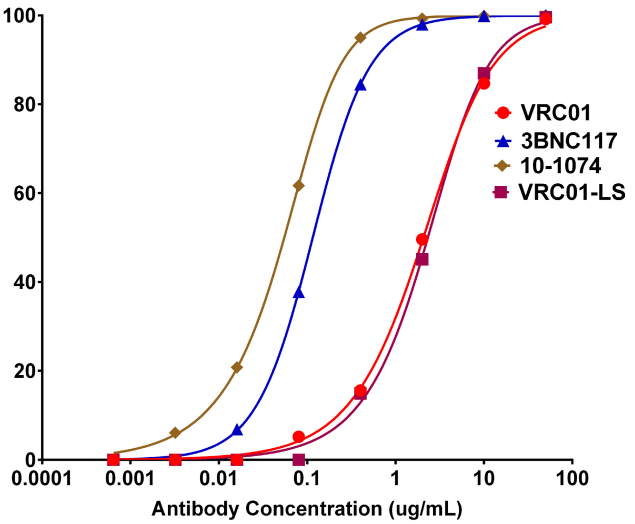

31. Endo, Y. *et al.* Short- and long-term clinical outcomes in rhesus monkeys inoculated with a highly pathogenic chimeric simian/human immunodeficiency virus. *J. Virol.* **74,** 6935–6945 (2000).
32. Li, M. *et al.* Human immunodeficiency virus type 1 env clones from acute and early subtype B infections for standardized assessments of vaccine-elicited neutralizing antibodies. *J. Virol.* **79,** 10108–10125 (2005).
33. Wei, X. *et al.* Emergence of resistant human immunodeficiency virus type 1 in patients receiving fusion inhibitor (T-20) monotherapy. *Antimicrob. Agents Chemother.* **46,** 1896–1905 (2002).

**a**



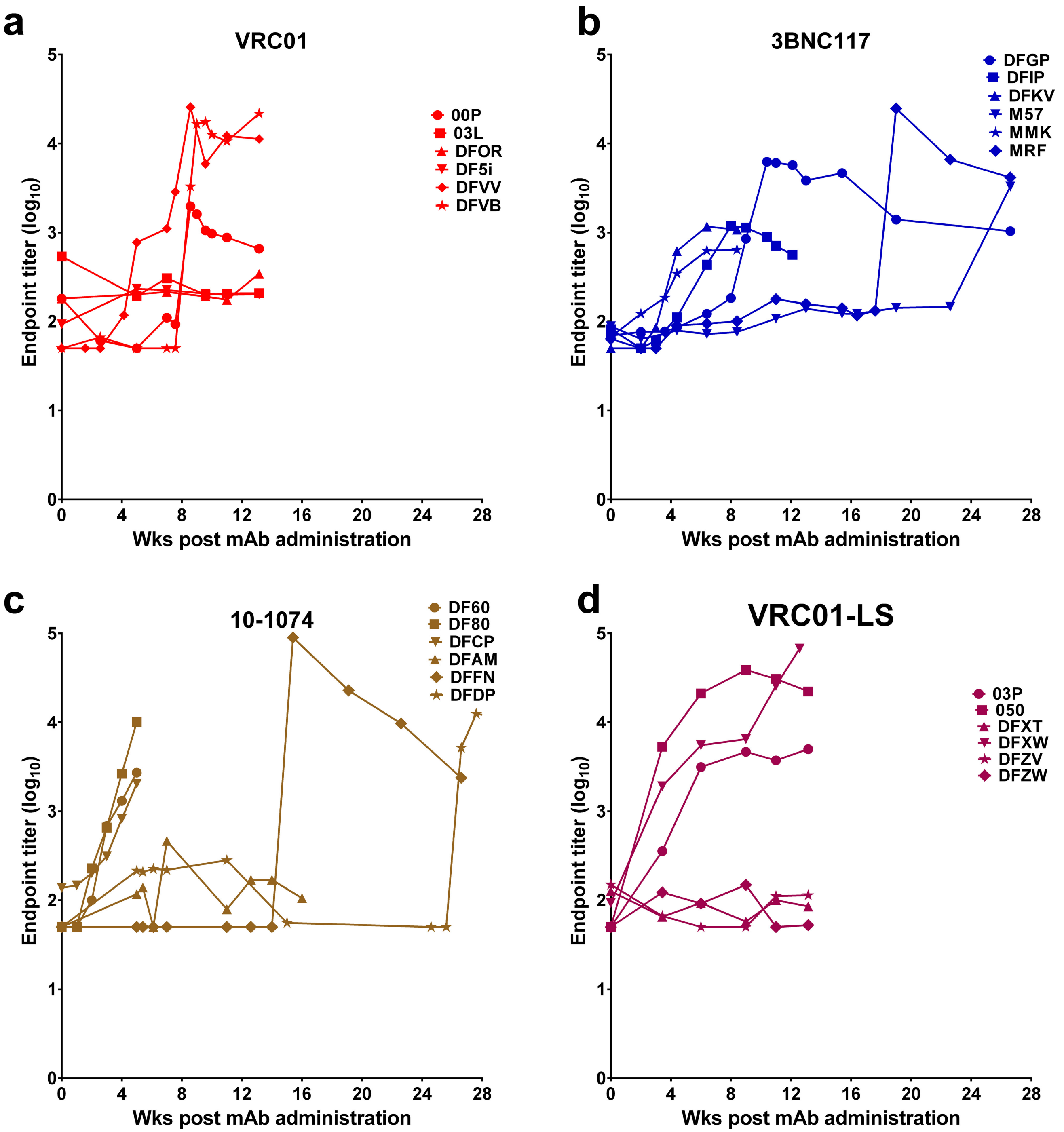

| | $IC_{50}$ (ug/mL) | $IC_{80}$ (ug/mL) |
|---|---|---|
| VRC01 | 0.67 | 2.04 |
| VRC01-LS | 0.58 | 1.76 |
| 3BNC117 | 0.06 | 0.19 |
| 10-1074 | 0.08 | 0.18 |

**b**

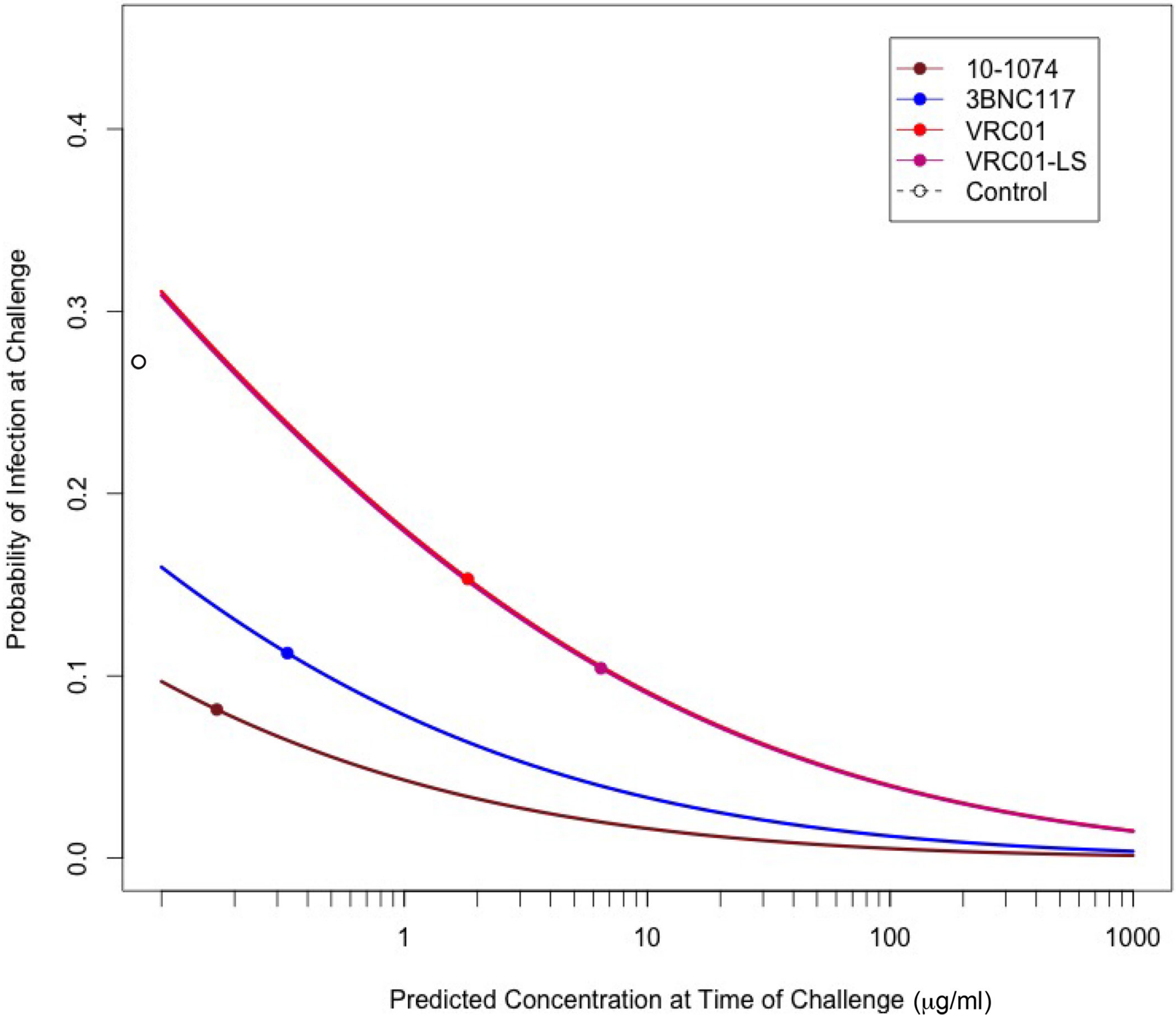



| | $IC_{50}$ (ug/mL) | $IC_{80}$ (ug/mL) |
|---|---|---|
| VRC01 | 2.06 | 7.14 |
| VRC01-LS | 2.32 | 6.83 |
| 3BNC117 | 0.12 | 0.32 |
| 10-1074 | 0.05 | 0.14 |

**Extended Data Figure 1 | Neutralization sensitivity of $SHIV_{AD8\text{-}EO}$ to four broadly acting neutralizing anti-HIV-1 MAbs. a**, Neutralizing activity of the indicated bNAbs was determined against $SHIV_{AD8\text{-}EO}$ pseudovirions using TZM-bl target cells. The calculated $IC_{50}$ and $IC_{80}$ values are shown at the bottom. **b**, Neutralizing activity of the indicated bNAbs was determined against replication competent $SHIV_{AD8\text{-}EO}$ in a single round TZM-bl infectivity assay. The calculated $IC_{50}$ and $IC_{80}$ values are shown at the bottom. The assay was performed in the presence of indinavir. Both experiments were performed twice.




**Extended Data Figure 2 | Development of anti-MAb immune responses in recipients of anti-HIV-1 bNAbs. a–d**, Longitudinal analysis of anti-VRC01, anti-3BNC117, anti-10-1074, and anti-VRC01-LS antibody responses, respectively, after a single intravenous infusion of indicated MAbs. This assay was performed twice.




**Extended Data Figure 3 | Predicted probability of infection as a function of antibody levels.** The per-challenge probability of infection was modelled as a function of antibody concentration at the time of each challenge using a probit regression model. The fitted probabilities from the models are plotted separately for each MAb group, with the estimated probability of infection for the control animals (0.27) indicated by the open circle adjacent to each ordinate. The VRC01 and VRC01-LS curves are superimposed. The points on each curve represent the median concentration at the time of breakthrough infection for each group of monkeys.

**Extended Data Table 1 | Plasma viral RNA and cell-associated viral RNA/DNA in rhesus macaques before breakthrough of infection**

| Animal | Wks post mAb treatment | Plasma Viral RNA (copies/ml) | SIV Gag RNA copies per $10^6$ cell eq | SIV Gag DNA copies per $10^6$ cell eq |
|---|---|---|---|---|
| DF60 | 7.4 | <2 | <1 | <1 |
| | 11.4 | <2 | <1 | <1 |
| DF80 | 3.6 | <2 | <1 | <1 |
| | 5.4 | <2 | <1 | <1 |
| DFAM | 11.4 | <2 | <1 | <1 |
| | 15.4 | <2 | <1 | <1 |
| DFCP | 5.4 | <2 | <1 | <1 |
| | 7.4 | <2 | <1 | <1 |
| DFDP | 13.6 | <2 | <1 | <1 |
| | 26.6* | 3,600,000 | 140000 | 610 |
| DFFN | 7.4 | <2 | <1 | <1 |
| | 11.4 | <2 | <1 | <1 |
| DFGP | 3.6 | <2 | <1 | <1 |
| | 5.4 | <2 | <1 | <1 |
| DFIP | 5.4 | <2 | <1 | <1 |
| | 6.4 | <2 | <1 | <1 |
| DFKV | 7.4 | <2 | <1 | <1 |
| | 11.4* | 10 | <1 | <1 |
| M57 | 11.4 | <2 | <1 | <1 |
| | 17.6 | <2 | <1 | <1 |
| MMK | 11.4 | <2 | <1 | <1 |
| | 15.4 | <2 | <1 | <1 |
| MRF | 9.4 | <2 | <1 | <1 |
| | 13.6 | <2 | <1 | <1 |
| 00P | 4.6 | <2 | <1 | <1 |
| | 6.6* | 270 | 3.2 | 4.2 |
| 03L | 8.6 | <2 | <1 | <1 |
| | 12.6* | 25,000 | 5100 | 25 |
| DF0R | 6.6 | <2 | <1 | <1 |
| | 10.6* | 73,000 | 16000 | 22 |
| DF5i | 6.6 | <2 | <1 | <1 |
| | 10.6 | <2 | <1 | <1 |
| DFVB | 4.6 | <2 | <1 | <1 |
| | 6.6* | 1,300 | <1 | <1 |
| DFVV | 2.6 | <2 | <1 | <1 |
| | 4.6* | 930 | <1 | <1 |
| 03P | 10.6 | <2 | <1 | <1 |
| | 14.6 | <2 | <1 | <1 |
| O50 | 4.6 | <2 | <1 | <1 |
| | 8.6 | <2 | <1 | <1 |
| DFXT | 10.6 | <2 | <1 | <1 |
| | 14.6 | <2 | <1 | <1 |
| DFXW | 10.6 | <2 | <1 | <1 |
| | 14.6 | <2 | <1 | <1 |
| DFZV | 10.6 | <2 | <1 | <1 |
| | 14.6 | <2 | <1 | <1 |
| DFZW | 10.6 | <2 | <1 | <1 |
| | 14.6* | 10 | <1 | <1 |

*Time point collected after breakthrough of infection. Ultrasensitive measurements of plasma SIV RNA or cell-associated SIV RNA and SIV DNA in peripheral blood mononuclear cells were determined by a nested, hybrid real-time/digital PCR assay.

**Extended Data Table 2 | VRC01 and 3BNC117 antibody concentrations in the plasma of macaques after a single administration of the indicated MAbs**

**VRC01 conc (μg/ml)**

| Wks | 00P | 03L | DFOR | DF5i | DFVB | DFVV |
|---|---|---|---|---|---|---|
| **0.0** | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 |
| **1.0** | 63 | 55.4 | 43.1 | 62.41 | 40.07 | 50.86 |
| **1.6** | 43.23 | 35.25 | 28.4 | 42.53 | 31.03 | 39.16 |
| **2.0** | 40.16 | 30.21 | 19.51 | 34.13 | 22.88 | 32.85 |
| **2.6** | 32.65 | 22.64 | 12.98 | 22.54 | 19.93 | 30.92 |
| **3.0** | 25.98 | 16.11 | 8.49 | 14 | 13.46 | 23.38 |
| **3.4** | 18.19 | 12.74 | 7.17 | 10.93 | 11.03 | 16.62 |
| **4.1** | 16.16 | 9.53 | 4.87 | 8.12 | 8.41 | 13.97 |
| **4.6** | 14.43 | 8.15 | 3.77 | 6.47 | 7.53 | 14.69 |
| **5.0** | 12.23 | 6.65 | 2.75 | 4.78 | 7 | 10.77 |
| **5.6** | 9.28 | 3.78 | 1.66 | 2.5 | 3.81 | 6.87 |
| **6.0** | 8.08 | 3.55 | 1.61 | 2.26 | 3 | 4.45 |
| **6.6** | 6.03 | 2.51 | 1.39 | 1.52 | 2.45 | 2.41 |
| **7.0** | 4.38 | 1.85 | 0.7 | 0.96 | 2.11 | 1.61 |
| **7.6** | 3.46 | 1.53 | 0.67 | 0.95 | 1.55 | 0.66 |
| **8.0** | 3.21 | 1.35 | 0.55 | 0.92 | 1.38 | |
| **8.6** | 1.06 | 1.32 | 0.5 | 0.78 | 0.8 | 0.22 |
| **9.0** | 0.83 | 1.12 | 0.39 | 0.51 | 0.52 | 0.1 |
| **9.6** | 0.33 | 0.89 | 0.32 | 0.48 | 0.5 | 0.1 |
| **10.0** | 0.27 | 0.78 | 0.31 | 0.45 | 0.37 | 0.1 |
| **10.6** | 0.1 | 0.73 | 0.3 | | 0.1 | 0.1 |
| **11.0** | 0.1 | 0.58 | 0.29 | 0.27 | 0.1 | 0.1 |
| **11.6** | 0.1 | 0.46 | 0.29 | 0.26 | 0.1 | 0.1 |
| **12.0** | 0.1 | 0.43 | 0.23 | 0.1 | 0.1 | 0.1 |
| **12.6** | 0.1 | 0.48 | 0.24 | 0.1 | 0.1 | 0.1 |
| **13.1** | 0.1 | 0.42 | 0.1 | 0.1 | 0.1 | 0.1 |
| **13.6** | | 0.60 | | | | |
| **14.0** | | 0.40 | | | | |
| **14.4** | | 0.10 | | | | |

**3BNC117 conc (μg/ml)**

| Wks | DFGP | DFIP | MMK | MRF | DFKV | M57 |
|---|---|---|---|---|---|---|
| **0.0** | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 |
| **1.0** | 72.6 | 78.1 | 88.7 | 64.3 | 65.5 | 63 |
| **1.4** | 57 | 41.4 | 65.7 | 49 | 44.2 | 41.9 |
| **2.0** | 47.4 | 39 | 54.6 | 35.1 | 30.8 | 37.9 |
| **3.0** | 19.6 | 29.2 | 38.8 | 29.1 | 23.2 | 23.4 |
| **3.6** | 14.4 | 20.9 | 32.9 | 17.2 | 14.7 | 12.7 |
| **4.0** | 12.2 | 16.8 | 20.7 | 14.7 | 8.6 | 11 |
| **4.4** | 8.6 | 15.5 | 20.3 | 11.3 | 4.1 | 8.8 |
| **5.4** | 7.8 | 8 | 12.6 | 9.1 | 0.4 | 4.9 |
| **6.4** | 2.5 | 4.4 | 9.3 | 5.7 | 0.1 | 3.3 |
| **7.4** | 1.5 | 1.4 | 4.3 | 3.3 | 0.1 | 2 |
| **8.0** | 1.3 | 0.5 | | | 0.1 | |
| **8.4** | 1.2 | 0.2 | 2.7 | 3 | 0.1 | 1.3 |
| **9.0** | 0.7 | 0.1 | | | | |
| **9.4** | 0.2 | 0.1 | 1.6 | 2.3 | 0.1 | 1.2 |
| **10.4** | 0.1 | | | | | |
| **11.0** | 0.1 | | 0.7 | 1.3 | 0.1 | 0.7 |
| **11.4** | 0.1 | | | | | |
| **12.1** | 0.1 | | | | | |
| **12.6** | | | 0.3 | 0.7 | 0.1 | 0.4 |
| **13.6** | | | 0.3 | 0.6 | 0.1 | 0.3 |
| **14.0** | | | 0.2 | 0.5 | | 0.3 |
| **14.6** | | | 0.2 | 0.4 | 0.1 | 0.2 |
| **15.0** | | | 0.1 | 0.4 | | 0.2 |
| **15.4** | | | 0.1 | 0.3 | | 0.2 |
| **16.0** | | | 0.1 | 0.3 | | 0.2 |
| **16.4** | | | 0.1 | 0.2 | | 0.2 |
| **17.0** | | | 0.1 | 0.2 | | |
| **17.6** | | | 0.1 | 0.2 | | 0.1 |
| **18.3** | | | 0.1 | 0.2 | | |
| **19.0** | | | 0.1 | 0.1 | | |

The plasma concentrations of the infused VRC01 and 3BNC117 were measured longitudinally in the indicated animals.

**Extended Data Table 3 | 10-1074 and VRC01-LS antibody concentrations in the plasma of macaques after a single administration of the indicated MAbs**

**10-1074 conc (µg/ml)**

| Wks | DFAM | DFCP | DF80 | DF60 | DFFN | DFDP |
|---|---|---|---|---|---|---|
| **0.0** | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 |
| **1.0** | 112.2 | 157.1 | 118.4 | 123.2 | 105 | 165.2 |
| **1.4** | 85.79 | 102.9 | 83.04 | 73.83 | 100.2 | 137.1 |
| **2.0** | 65.71 | 75.66 | 54.04 | 51.46 | 70.47 | 121.3 |
| **3.0** | 40.26 | 30.14 | 19.31 | 13.14 | 34.16 | 115.3 |
| **3.6** | 38.54 | 18.54 | 3.09 | 2.92 | 33.9 | 76.13 |
| **4.0** | 29.05 | 9.41 | 0.48 | 0.82 | 26.9 | 66.49 |
| **4.4** | 23.98 | 4.13 | 0.1 | 0.27 | 23.48 | 58.02 |
| **5.0** | 19.22 | 1.07 | 0.1 | 0.1 | 14.98 | 48.57 |
| **5.4** | 15.36 | 0.34 | 0.1 | 0.1 | 11.37 | 50.83 |
| **6.1** | 10.25 | 0.1 | 0.1 | 0.1 | 7.47 | 43.68 |
| **6.4** | 11.31 | 0.1 | 0.1 | 0.1 | 7.15 | 53.86 |
| **7.0** | 11.17 | 0.1 | 0.1 | 0.1 | 5.95 | 34.49 |
| **7.4** | 10.01 | 0.1 | 0.1 | 0.1 | 6.78 | 22.26 |
| **8.0** | 8.04 | 0.1 | 0.1 | 0.1 | 3.19 | 18.74 |
| **8.4** | 6.29 | 0.1 | 0.1 | 0.1 | 2.74 | 16.32 |
| **9.0** | 5.69 | 0.1 | 0.1 | 0.1 | 2.43 | 14.52 |
| **9.4** | 4.52 | 0.1 | 0.1 | 0.1 | 1.65 | 20.28 |
| **10.4** | 3.78 | 0.1 | 0.1 | 0.1 | 1.05 | 13.05 |
| **11.0** | 2.51 | 0.1 | 0.1 | 0.1 | 0.71 | 9.75 |
| **11.4** | 2.21 | 0.1 | 0.1 | 0.1 | 0.84 | 8.53 |
| **12.1** | 1.71 | 0.1 | 0.1 | 0.1 | 0.65 | 6.48 |
| **12.6** | 0.85 | 0.1 | 0.1 | 0.1 | 0.37 | 3.94 |
| **13.3** | 0.72 | 0.1 | | 0.1 | 0.27 | 5.44 |
| **13.6** | 1.3 | | 0.1 | 0.1 | 0.34 | 4.5 |
| **14.0** | 0.91 | | | 0.1 | 0.24 | 4.42 |
| **14.6** | 0.72 | 0.1 | 0.1 | 0.1 | 0.19 | 4.19 |
| **15.0** | 0.61 | | | | 0.1 | 4.21 |
| **15.4** | 0.58 | 0.1 | 0.1 | 0.1 | 0.1 | 3.5 |
| **16.0** | 0.45 | | | | | 3.72 |
| **17.6** | | | | | | 1.87 |
| **18.3** | | | | | | 1.95 |
| **19.0** | | | | | 0.1 | 1.8 |
| **19.6** | | | | | | 1.6 |
| **20.0** | | | | | | 1.09 |
| **20.6** | | | | | | 1.03 |
| **21.0** | | | | | | 0.93 |
| **22.0** | | | | | | 0.78 |
| **22.6** | | | | | 0.1 | 0.73 |
| **23.0** | | | | | | 0.62 |
| **23.4** | | | | | | 0.59 |
| **24.1** | | | | | | 0.46 |
| **24.6** | | | | | | 0.41 |
| **25.0** | | | | | | 0.43 |
| **25.6** | | | | | | 0.3 |
| **26.6** | | | | | 0.1 | 0.1 |

**VRC01-LS conc (µg/ml)**

| Wks | 03P | 050 | DFXT | DFXW | DFZV | DFZW |
|---|---|---|---|---|---|---|
| **0.0** | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 |
| **1.0** | 225.6 | 192.3 | 191.8 | 226.3 | 206 | 234 |
| **1.6** | 169.9 | 181.9 | 153.7 | 205.5 | 185.8 | 169.8 |
| **2.0** | 147.1 | 183.8 | 144.9 | 192.1 | 196.5 | 180.9 |
| **2.6** | 161.1 | 158.4 | 134.1 | 175.7 | 143.3 | 160 |
| **3.0** | 139.6 | 121 | 114.2 | 146.9 | 122.4 | 126.1 |
| **3.4** | 100.8 | 123.2 | 97.5 | 149.2 | 99.62 | 109.9 |
| **4.1** | 85.8 | 99.2 | 85.39 | 115.4 | 83.86 | 110.5 |
| **4.6** | 80.75 | 81.47 | 84.05 | 96.66 | 90.35 | 89.88 |
| **5.0** | 74.15 | 70.61 | 81.56 | 94 | 76.03 | 78.98 |
| **5.6** | 66.4 | 50.01 | 62.09 | 66.09 | 64.61 | 57.82 |
| **6.0** | 68.3 | 46.59 | 68.85 | 59.62 | 64.89 | 55.32 |
| **6.6** | 48.25 | 32.37 | 48.55 | 49.35 | 59.46 | 52.37 |
| **7.0** | 41.47 | 19.19 | 42.24 | 42.01 | 45.4 | 43.91 |
| **7.6** | 35.11 | 12.16 | 35.04 | 35.43 | 37.21 | 33.66 |
| **8.0** | 35.83 | 9.45 | 44.51 | 32.74 | 46.14 | 40.3 |
| **8.6** | 28.34 | 5.09 | 33.09 | 28.2 | 37.1 | 34.81 |
| **9.0** | 27.42 | 3.01 | 29.52 | 25.81 | 31.81 | 29.42 |
| **9.6** | 20.82 | 1.5 | 27.67 | 20.34 | 26.1 | 21.27 |
| **10.0** | 19.85 | 0.2 | 32.19 | 17.53 | 27.91 | 27.6 |
| **10.6** | 15.94 | 0.2 | 20.39 | 9.9 | 24.64 | 21.42 |
| **11.0** | 13.36 | 0.2 | 19.64 | 5.28 | 20.15 | 19.54 |
| **11.6** | 12.62 | 0.2 | 17.85 | 2.06 | 17.7 | 16.72 |
| **12.0** | 12.52 | 0.2 | 16.9 | 0.64 | 18.82 | 18.1 |
| **12.6** | 12.4 | 0.2 | 15.67 | 0.2 | 15.06 | 15.68 |
| **13.1** | 9.66 | 0.2 | 14.02 | 0.2 | 13.15 | 13.11 |
| **13.6** | 10.40 | | 15.20 | | 16.80 | 14.10 |
| **14.0** | 9.70 | | 14.10 | | 16.20 | 13.60 |
| **14.4** | 8.10 | | 12.70 | | 13.50 | 11.90 |
| **15.0** | 8.00 | | 12.00 | | 14.40 | 11.80 |
| **15.6** | 7.00 | | 11.00 | | 12.80 | 10.00 |
| **16.0** | 6.00 | | 10.40 | | 12.70 | 9.20 |
| **16.6** | 4.40 | | 9.00 | | 11.10 | 6.30 |
| **17.0** | 4.60 | | 9.30 | | 10.90 | 2.00 |
| **17.6** | 4.10 | | 7.50 | | 10.00 | 2.40 |
| **18.1** | 2.40 | | 6.60 | | 8.20 | 1.70 |
| **18.6** | 1.95 | | 5.76 | | 7.52 | 1.29 |
| **19.0** | | | 5.64 | | 6.57 | |
| **19.6** | 1.80 | | 4.70 | | 2.98 | 0.84 |
| **20.0** | | | 3.29 | | | |
| **20.6** | 1.07 | | 2.06 | | 2.10 | 0.55 |
| **23.0** | 0.10 | | 1.33 | | 0.10 | 0.30 |
| **26.3** | 0.10 | | 0.32 | | 0.10 | 0.12 |

The plasma concentrations of the infused 10-1074 and VRC01-LS were measured longitudinally in the indicated animals.

doi:10.1038/nature17947

# EBI2 augments Tfh cell fate by promoting interaction with IL-2-quenching dendritic cells

Jianhua Li[1,2,3], Erick Lu[1,2], Tangsheng Yi[1,2]† & Jason G. Cyster[1,2]

**T follicular helper (Tfh) cells are a subset of T cells carrying the CD4 antigen; they are important in supporting plasma cell and germinal centre responses[1,2]. The initial induction of Tfh cell properties occurs within the first few days after activation by antigen recognition on dendritic cells, although how dendritic cells promote this cell-fate decision is not fully understood[1,2]. Moreover, although Tfh cells are uniquely defined by expression of the follicle-homing receptor CXCR5 (refs 1, 2), the guidance receptor promoting the earlier localization of activated T cells at the interface of the B-cell follicle and T zone has been unclear[3–5]. Here we show that the G-protein-coupled receptor EBI2 (GPR183) and its ligand 7α,25-dihydroxycholesterol mediate positioning of activated CD4 T cells at the interface of the follicle and T zone. In this location they interact with activated dendritic cells and are exposed to Tfh-cell-promoting inducible co-stimulator (ICOS) ligand. Interleukin-2 (IL-2) is a cytokine that has multiple influences on T-cell fate, including negative regulation of Tfh cell differentiation[6–10]. We demonstrate that activated dendritic cells in the outer T zone further augment Tfh cell differentiation by producing membrane and soluble forms of CD25, the IL-2 receptor α-chain, and quenching T-cell-derived IL-2. Mice lacking EBI2 in T cells or CD25 in dendritic cells have reduced Tfh cells and mount defective T-cell-dependent plasma cell and germinal centre responses. These findings demonstrate that distinct niches within the lymphoid organ T zone support distinct cell fate decisions, and they establish a function for dendritic-cell-derived CD25 in controlling IL-2 availability and T-cell differentiation.**

EBI2 is expressed by CD4 T cells[11–14], but whether it has a role in positioning T cells during the early stages of activation has been unclear. Using an ovalbumin (OVA)-specific T-cell antigen receptor (TCR) transgenic (OTII) system involving transfer of OTII T cells to wild-type (WT) hosts, we found that EBI2 was upregulated on cognate splenic T cells within 12 h of immunization with a particulate form of OVA (sheep red blood cell (SRBC) conjugated), and it remained high at day 2 (Extended Data Fig. 1a). Similar EBI2 induction occurred after immunization with OVA in lipopolysaccharide, on lymph node T cells after immunization with OVA in alum, and *in vitro* after T-cell activation by anti-CD3 and -CD28 (Extended Data Fig. 1b–e). Migration to 7α,25-dihydroxycholesterol (7α,25-OHC) was augmented at these time points (Extended Data Fig. 1f). Analysis of spleen sections showed that transferred WT T cells accumulated in the outer T zone at 12 h and day 1 of the SRBC-OVA response and they remained enriched in this location at day 2 (Fig. 1a). EBI2 knockout (KO) T cells, by contrast, failed to accumulate in the outer T zone at either time point and instead remained dispersed throughout the T zone (Fig. 1a). Quantitative analysis using a mixed transfer system confirmed that the activated EBI2 KO cells had less access than control cells to the outer T zone (Fig. 1b and Extended Data Fig. 1g). Similar findings were made at day 2 after immunization with OVA-expressing *Listeria monocytogenes* (Fig. 1c) and with OVA in lipopolysaccharide (Extended Data Fig. 1h). WT OTII T cells also moved to the B–T zone interface in lymph nodes after immunization with alum-OVA, but EBI2-deficient T cells failed to re-localize (Fig. 1d and Extended Data Fig. 1i). Activated T-cell positioning in the outer T zone was directed by 7α,25-OHC as it was dependent on the enzymes needed for its synthesis (Cyp7b1 and Ch25h) and catabolism (Hsd3b7) (Extended Data Fig. 1j).

Flow cytometric analysis for the early activation marker CD69 showed that co-transferred EBI2 KO and WT T cells were comparably activated at day 2 of the SRBC-OVA response (Fig. 2a), indicating similar initial exposure to cognate MHC class II-peptide complexes.

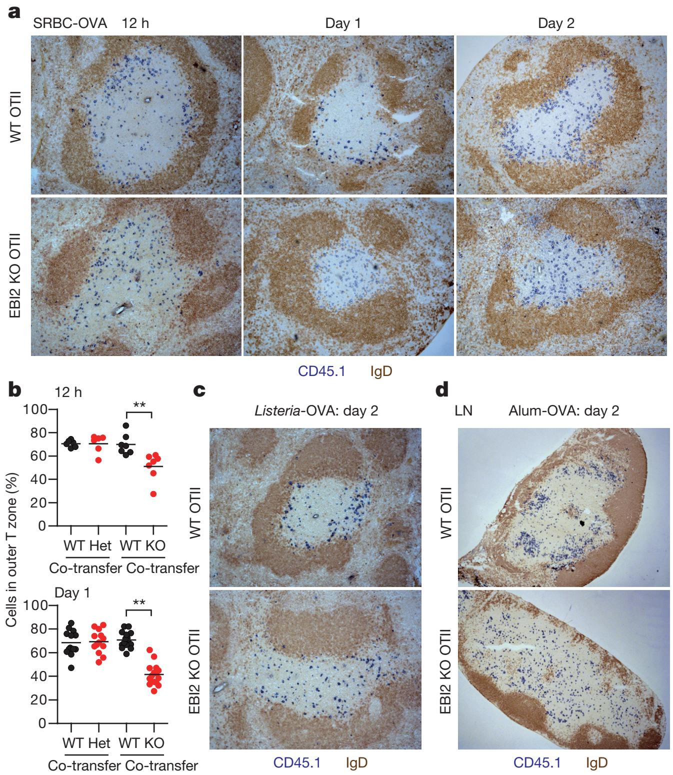



**Figure 1 | EBI2 promotes positioning of newly activated CD4 T cells in the outer T zone. a**, Immunohistochemical analysis of spleens for transferred WT or EBI2 KO OTII CD45.1$^{+}$ T cells (blue) and endogenous B cells (IgD, brown) at 12 h, 1 day and 2 days after SRBC-OVA immunization. **b**, Fraction of WT and EBI2 het or KO OTII T cells in the outer quarter of the splenic T zone at 12 h and 1 day after SRBC-OVA. Sections were stained as in Extended Data Fig. 1g. See Methods for details. **c**, **d**, As for a except mice were immunized with *Listeria*-OVA (**c**) or alum-OVA and inguinal lymph nodes were analysed (**d**). $^{**}P < 0.01$ by Student's *t*-test. Data are representative of three (**a**, **b**) or two (**c**–**e**) experiments with at least three (**a**) or two (**b**–**e**) mice per group.

[1]Department of Microbiology and Immunology, University of California, San Francisco, San Francisco, California 94143, USA. [2]Howard Hughes Medical Institute, University of California, San Francisco, San Francisco, California 94143, USA. [3]Key Laboratory of Medical Molecular Virology, Department of Medical Microbiology, School of Basic Medical Sciences, Shanghai Medical College, Fudan University, Shanghai, China. †Present address: Department of Discovery Immunology, Genentech, South San Francisco, California 94080, USA.

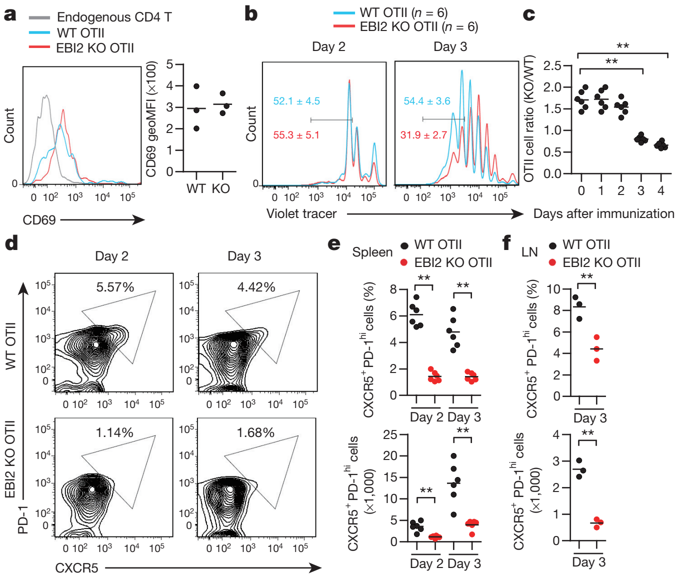



**Figure 2 | Defective differentiation of EBI2-deficient T cells to follicular helpers. a**, CD69 expression on WT OTII, EBI2 KO OTII and endogenous CD4 T cells in transfer recipient spleens 2 days after SRBC-OVA immunization. Histograms show representative FACS and graphs show summary geoMFI data for three mice of each type. **b**, Proliferation of co-transferred WT and EBI2 KO OTII T cells monitored by violet tracer dye dilution at days 2 and 3 after immunization. Numbers indicate mean percentage (±s.d.) of cells in the indicated gate ($n = 6$). **c**, Summary of data from **b** shown as a ratio of KO/WT OTII T-cell number at the indicated days. **d**, Flow cytometric analysis of co-transferred WT and EBI2 KO OTII T cells for PD-1 and CXCR5 at days 2 and 3 after immunization. Numbers indicate frequency of cells in gated region. **e**, Summary of data of the type in **d**: frequency (top) and number (bottom) of $CXCR5^{+}PD\text{-}1^{hi}$ $CD4^{+}$ OTII T cells. **f**, Frequency and number of $CXCR5^{+}PD\text{-}1^{hi}$ $CD4^{+}$ OTII T cells in peripheral lymph nodes of mice of the type in **d**, immunized with alum-OVA. $*P < 0.05$ and $**P < 0.01$ by analysis of variance (ANOVA) (**d**, **f**) or Student's *t*-test (**g**). Data are representative of three (**a**–**f**) or two (**g**) experiments with at least three mice per group.

Upregulation of the co-stimulatory molecules ICOS and OX40 also occurred to an equivalent extent (Extended Data Fig. 2a). Proliferation began by day 2 and at this time point the WT and EBI2 KO cells responded similarly (Fig. 2b, c). However, by day 3, the EBI2-deficient cells were undergoing less proliferation and their numbers increased more slowly (Fig. 2b, c). This was not due to a direct effect of 7α,25-OHC on T-cell proliferation (Extended Data Fig. 2b, c). Tracking of differentiation markers on the *in vivo* activated T cells revealed that EBI2 KO cells were compromised in their induction of a Tfh cell phenotype, as assessed by CXCR5, PD-1 (Fig. 2d, e), Bcl6, and *Il21* expression (Extended Data Fig. 2d–f). EBI2-deficient OTII T cells also differentiated less efficiently into Tfh cells in lymph nodes (Fig. 2f). We also observed reduced Tfh cell responses to *Listeria*-OVA, reduced polyclonal EBI2 KO Tfh cell responses to SRBCs and reduced germinal centre and plasma cell responses to these antigens (Supplementary Information and Extended Data Fig. 3a–j).

Tfh cell differentiation is promoted by interaction both with dendritic cells (DCs) and with B cells, and time-course studies indicate that DCs are critical early while B cells play a later role[1,2]. Consistent with these requirements, WT OTII T cells showed a partial reduction in Tfh cell differentiation at day 3 of the response in MD4 immunoglobulin (Ig)-transgenic mice lacking cognate B cells capable of OVA antigen presentation to OTII T cells (Fig. 3a). Importantly, however, EBI2 KO OTII T cells formed fewer Tfh cells than WT OTII T cells in the MD4 recipients, indicating that T-cell EBI2 expression augmented Tfh cell development at early time points in a B-cell-independent manner.

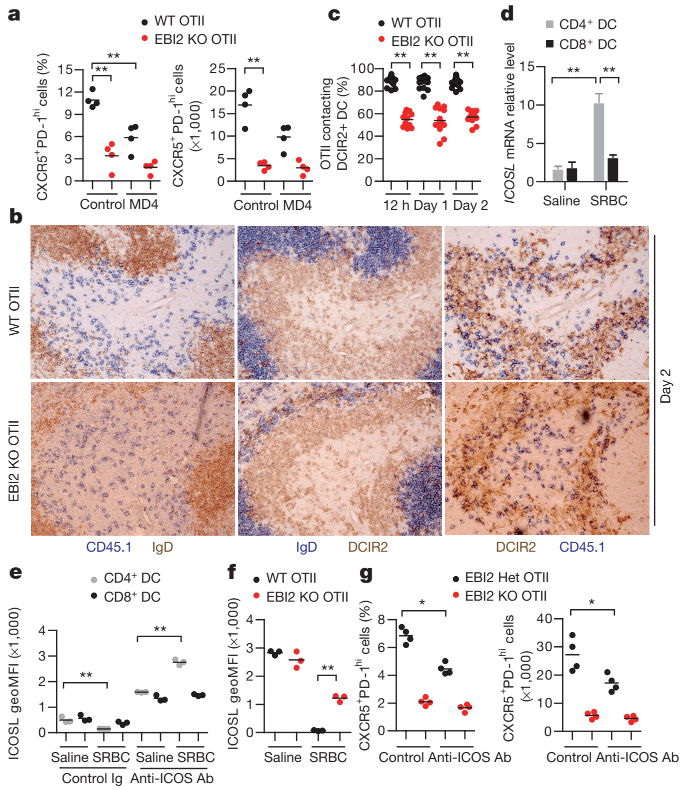



**Figure 3 | T-cell EBI2 is required for $CD4^{+}$ DC-mediated augmentation of Tfh cell induction. a**, Frequency and number of $CXCR5^{+}PD\text{-}1^{hi}$ WT and EBI2 KO OTII T cells in control or MD4 recipient spleens at day 3 after immunization with SRBC-OVA. **b**, Immunohistochemical analysis of consecutive sections from WT transfer recipient spleens at day 2 after immunization, stained for (left) WT or EBI2 KO OTII $CD45.1^{+}$ T cells (blue) and B cells (IgD, brown), (centre) $DCIR2^{+}$ DCs (brown) and B cells (IgD, blue) and (right) OTII $CD45.1^{+}$ T cells (blue) and $DCIR2^{+}$ DCs (brown). **c**, Frequency of WT or EBI2 KO OTII T cells contacting $DCIR2^{+}$ DCs determined in sections of the type in **b** and Extended Data Fig. 4e, at the indicated times after immunization. **d**, *Icosl* mRNA abundance in splenic $CD4^{+}$ and $CD8^{+}$ $CD11c^{+}$ DCs from mice immunized 12 h earlier with saline or SRBCs, shown relative to the saline control. **e**, Summary data of ICOSL surface levels for DCs of the type in **d** from mice also treated with control or ICOS blocking antibody. **f**, ICOSL surface levels on $CD4^{+}$ splenic DCs from mice that had received WT or EBI2 KO OTII T cells, 12 h after immunization with SRBC-OVA. Recipient mice were CD28 KO. See Supplementary Information for details. **g**, Frequency and number $CXCR5^{+}PD\text{-}1^{hi}$ WT and EBI2 KO OTII T cells in spleens from mice treated with ICOS blocking antibody, analysed at day 3 after immunization. $*P < 0.05$ and $**P < 0.01$ by ANOVA (**a**–**d**, **e**) or Student's *t*-test (**f**, **g**). Data are representative of three (**a**–**c**) or two (**d**–**g**) experiments with at least three (**a**, **c**–**g**) or two (**b**) mice per group (error bars (**d**), s.e.m.).

Similar observations were made in B-cell-deficient μMT mice (Extended Data Fig. 4a). These findings led us to test whether EBI2 was required in T cells for some type of interaction with DCs. Ablation of DCs using Zbtb46-diphtheria toxin receptor (DTR) mice[15] caused a complete block in Tfh cell generation (Extended Data Fig. 4b, c), consistent with previous studies using other DC ablation approaches[16]. Splenic CD4 and DCIR2 co-expressing DCs re-localize from bridging channels to the outer T zone within 6 h of immunization with SRBCs[17,18], and the cells remain in this region for at least 2 days (Fig. 3b and Extended Data Fig. 4d). At 12 h, day 1 and day 2 time points, almost the entire population of activated WT OTII T cells co-localized with the activated $DCIR2^{+}$ DCs (Fig. 3b, c and Extended Data Fig. 4e). By contrast, activated EBI2 KO T cells were broadly distributed and only partly overlapped with the $DCIR2^{+}$ DCs (Fig. 3b, c and Extended Data Fig. 4e). Using mice

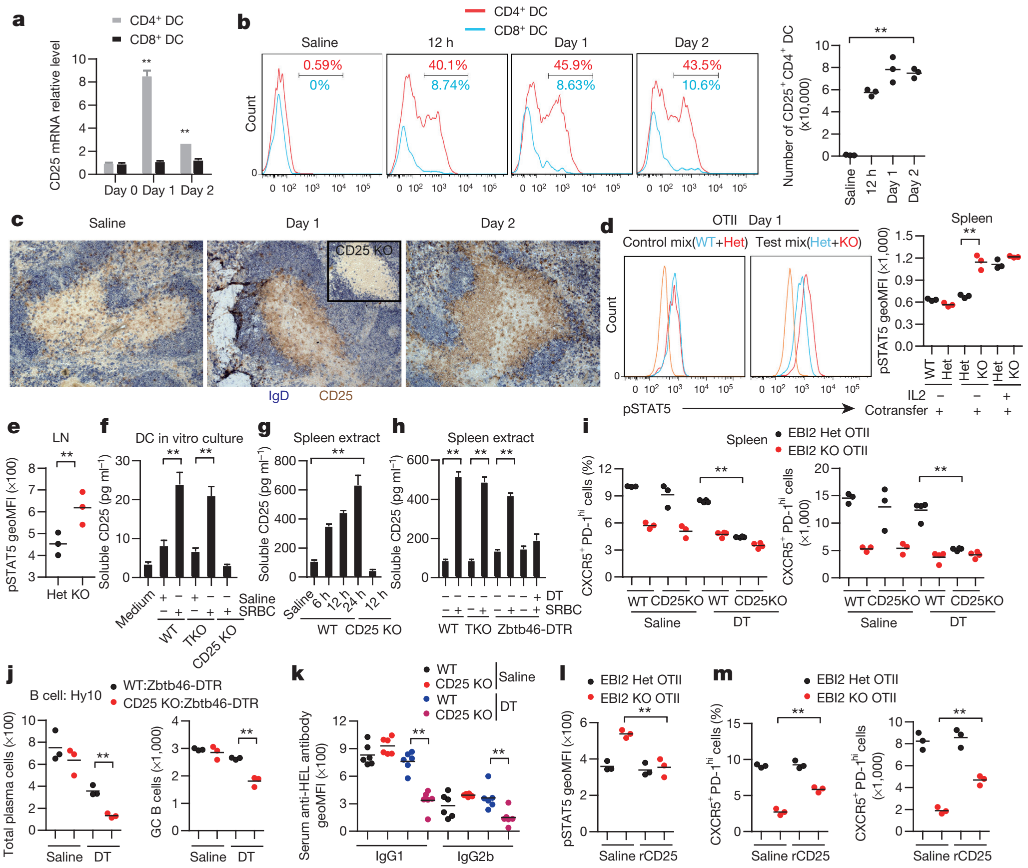



**Figure 4 | DC CD25 expression reduces IL-2 signalling in activated CD4 T cells, favouring their differentiation to follicular helpers.** **a**, **b**, CD25 transcript (**a**) and surface (**b**) levels on CD4$^+$ and CD8$^+$ splenic DCs from mice immunized with saline or with SRBCs 12 h, 1 or 2 days earlier. Transcript levels are plotted relative to the day 0 mean for each DC type. **c**, Immunohistochemical analysis of spleens from WT mice immunized with saline or SRBCs, stained to detect IgD (blue) and CD25 (brown). Inset shows 12 h SRBC immunized CD25 KO. **d**, Flow cytometry of pSTAT5 in WT, EBI2 Het or EBI2 KO OTII T cells in mice that received mixtures of CD45.2 and CD45.1/2 marked cells, stained *ex vivo* 1 day after SRBC-OVA immunization. Left: example histogram plots of gated OTII T cells. Right: summary geoMFI data for three mice of each type in one experiment, including mice injected with IL-2 as a positive control. Orange histogram indicates endogenous CD4$^+$ T cells. **e**, Summary of pSTAT5 levels in control (Het) and EBI2 KO OTII T cells in lymph nodes at day 3 after alum-OVA immunization. **f**, Soluble CD25 detected by ELISA in culture supernatants of splenic CD4$^+$ DCs from WT, TCR KO or CD25 KO mice immunized 1 day earlier with saline or SRBCs, or medium alone. **g**, **h**, Soluble CD25 detected by ELISA in spleen extracts taken from WT, CD25 KO, TCR KO, or Zbtb46-DTR treated with DT, immunized as indicated (**g**) or 12 h (**h**) before analysis. **i**, Frequency and number of CXCR5$^+$PD-1$^{hi}$ control (EBI2 Het) and EBI2 KO OTII T cells in spleens from WT:Zbtb46-DTR or CD25 KO:Zbtb46-DTR mixed BM chimeras pre-treated with saline or DT, at day 3 after SRBC-OVA. **j**, HEL-binding plasma cell and germinal centre B cell numbers in spleens from WT:Zbtb46-DTR (control) or CD25 KO:Zbtb46-DTR chimaeras treated with saline or DT and transferred with Hy10 B cells, at day 5 after immunization with HEL-SRBC. **k**, Serum IgG1 and IgG2b anti-HEL antibody in mice of the type in j analysed by FACS of HEL-conjugated mouse RBCs. **l**, **m**, Summary geoMFI of pSTAT5 levels (**l**) and CXCR5$^+$PD-1$^{hi}$ cell frequencies and numbers of control (Het) and EBI2 KO OTII T cells in WT mice at day 1 (**l**) and day 3 (**m**) after SRBC-OVA immunization and treatment with saline or recombinant CD25 (rCD25). TKO, TCRβδ KO. *$P < 0.05$ and **$P < 0.01$ by ANOVA (**a**, **b**, **f**–**h**, **k**) or Student's *t*-test (**d**, **e**, **i**, **j**, **l**). Data are representative of three (**a**, **b**, **d**) or two (**c**, **e**–**m**) experiments with at least three (**a**, **b**, **d**–**f**, **h**–**m**) or two (**c**, **g**) mice per group (error bars (**a**, **f**–**h**), s.e.m.).

with deficiencies in DCs, we established that CD4$^+$ but not CD8$^+$ DCs were important for Tfh cell induction by SRBC-OVA (Supplementary Information and Extended Data Fig. 4f–k).

ICOS signalling is important for Tfh cell differentiation[1,2] and splenic DCs upregulated ICOS ligand (ICOSL) mRNA after activation by SRBCs, with the extent of upregulation being greater in CD4$^+$ than CD8$^+$ DCs (Fig. 3d). However, flow cytometric analysis showed CD4$^+$ DCs activated for 12 h had low surface ICOSL staining (Fig. 3e and Extended Data Fig. 4l). Since ICOSL undergoes rapid ectodomain shedding after ICOS engagement[19,20], we considered the possibility that surface levels were reduced because of interactions with ICOS-high activated T cells. Consistent with this idea, when SRBC-immunized mice were also treated with an ICOS blocking antibody to prevent ICOS-induced ICOSL shedding[19,20], CD4$^+$ but not CD8$^+$ DCs showed increased ICOSL surface abundance compared with unimmunized mice (Fig. 3e and Extended Data Fig. 4l). We took advantage of the sensitivity of ICOSL to ICOS-induced shedding as a method to measure the amount of interaction between DCs and cognate T cells. Twelve hours after SRBC-OVA immunization, ICOSL levels were higher on CD4$^+$ DCs in mice harbouring EBI2-deficient OTII T cells than WT OTII T cells (Fig. 3f). These data suggest that the reduced Tfh differentiation of EBI2 KO OTII T cells occurs at least in part because of lower ICOS engagement with ICOSL on CD4$^+$ DCs. However, in mice treated with an ICOS blocking antibody, although Tfh differentiation of control OTII T cells was reduced, EBI2-deficient OTII T cells were still more defective (Fig. 3g), indicating an ICOS-independent influence of EBI2 in augmenting Tfh cell fate. An assessment of mRNA levels of other factors established to have an effect on Tfh differentiation (IL-6, transforming growth factor-β (TGF-β)) showed that they were similarly expressed in CD4$^+$ and CD8$^+$ DCs and they were therefore not considered likely factors accounting for the EBI2-dependence of Tfh cell differentiation (Extended Data Fig. 4n).

To search for surface or secreted DC-derived factors that might augment Tfh cell differentiation, we performed RNA sequencing (RNA-seq) analysis on $CD4^+$ DCs from the spleens of saline or SRBC immunized mice. This analysis revealed CD25, the high affinity IL-2 receptor $\alpha$-chain, as one of the most strongly induced genes in SRBC-activated DCs (Extended Data Fig. 5a). CD25 mRNA induction occurred rapidly after immunization and remained elevated for at least 2 days (Fig. 4a), and many $CD4^+$ DCs were surface positive for CD25 over this time frame (Fig. 4b). $CD8^+$ DCs showed little induction of CD25 mRNA or protein under these immunization conditions (Fig. 4a, b). Analysis of lymph node DCs after OVA plus alum immunization revealed upregulation of CD25 on migratory $CD11b^+$ DCs at day 1 and 2 (Extended Data Fig. 5b). Staining of spleen sections identified $CD25^+$ cells in the unstimulated T zone that are likely $CD25^{hi}$ regulatory T cells, but also showed broad induction of CD25 in the outer T zone within 12 h of SRBC immunization in a pattern resembling the $DCIR2^+$ DC distribution (Fig. 4c). Similar appearance of CD25 staining in the T zone was seen in T-cell-deficient mice, providing evidence that expression by activated DCs was being detected (Extended Data Fig. 5c). CD25 needs to associate with CD122 (IL-2R$\beta$) and IL-2R$\gamma$ to transmit signals in response to IL-2 (ref. 10). Despite expression of CD25, activated $CD4^+$ DCs showed minimal CD122 mRNA and protein expression and they did not respond to IL-2 as assessed by intracellular pSTAT5 staining (Extended Data Fig. 5d–f). These data led us to consider the possibility that activated DCs express CD25 to alter IL-2 availability in the outer T zone.

IL-2 has pleiotropic roles in directing T-cell fate, including a negative influence on Tfh cell differentiation[6–10]. Despite equivalent IL-2 production and receptor expression (Supplementary Information and Extended Data Fig. 5g–i), EBI2 KO T cells showed more IL-2R signalling than WT T cells at day 1 after immunization, as evidenced by higher pSTAT5 levels (Fig. 4d), suggesting that the EBI2 KO T cells were being exposed to more IL-2. The elevated induction of pSTAT5 in EBI2 KO T cells was also seen in lymph nodes after immunization with alum-OVA (Fig. 4e). Blimp1, encoded by *Prdm1*, is induced by IL-2 and negatively regulates expression of Bcl6, a factor essential for Tfh cell development[2]. In agreement with higher IL-2 exposure, the EBI2 KO T cells showed greater *Prdm1* expression at day 3 (Extended Data Fig. 5j).

The above findings led us to test whether DCs antagonize IL-2 availability to activated T cells in the outer T zone. Consistent with this possibility, when *in vivo* activated DCs were incubated briefly *in vitro* they were found to release soluble CD25 (sCD25) into the culture supernatant (Fig. 4f). DC production of sCD25 was not dependent on interaction with T cells (Fig. 4f). Analysis of spleen tissue extracts showed elevated sCD25 production at 6 h after SRBC immunization and this was maintained at 24 h (Fig. 4g) and occurred in a T-cell-independent and DC-dependent manner (Fig. 4h). To determine whether the sCD25 functioned as an IL-2 antagonist, supernatants from cultures of activated $CD4^+$ DCs were tested in a bioassay for their ability to inhibit IL-2R signalling. Regulatory T cells were used as the reporter cells in this bioassay since they were more sensitive to low-dose IL-2 than Tfh cells (J.L. and J.G.C., unpublished observations). Supernatants from cultured SRBC-activated WT but not CD25 KO $CD4^+$ DCs were able to antagonize IL-2R signalling in T cells (Extended Data Fig. 5l).

To determine whether DCs were regulating T-cell differentiation *in vivo* by production of CD25, we generated bone marrow (BM) chimaeric mice that lacked CD25 in DCs (Supplementary Information and Extended Data Fig. 6a–e). In these recipients, control (EBI2 Het) OTII T cells were compromised in their ability to take on a Tfh cell fate whereas EBI2 KO T cells were only mildly affected (Fig. 4i). Staining of tissue sections from the DT-treated BM chimaeras established that the high CD25 abundance in the outer T zone was dependent on CD25 expression by DCs (Extended Data Fig. 6f). Splenic Tfh cell responses to *Listeria*-OVA and lymph node Tfh cell responses to alum-OVA were also diminished in mice lacking CD25 on DCs (Extended Data Fig. 6g, h). Moreover, mice lacking CD25 in DCs and harbouring transgenic B cells specific for hen egg lysozyme (HEL) mounted reduced plasma cell and germinal centre responses to HEL-conjugated SRBCs (Fig. 4j and Extended Data Fig. 6i, j), which was associated with reduced serum anti-HEL IgG1 and IgG2b antibody levels (Fig. 4k). To further test whether reduced exposure of EBI2 KO T cells to sCD25 could account for the defective Tfh cell induction, we treated mice with recombinant sCD25 at day 0 and 1 of the OTII T-cell response to SRBC-OVA. This treatment was sufficient to elevate sCD25 levels in tissue extracts (Extended Data Fig. 6k), to antagonize pSTAT5 over-induction in the EBI2 KO T cells (Fig. 4l) and to partly rescue Tfh cell differentiation (Fig. 4m). A full restoration of the Tfh response was not expected given that EBI2-deficient cells are also compromised in accessing ICOSL and possibly other Tfh-cell-promoting signals from cells in the outer T zone.

This work establishes a role for EBI2 and 7$\alpha$,25-OHC in positioning activated T cells at the follicle–T-zone interface, promoting contact with Tfh cell-priming $ICOSL^{hi}$ $CD25^+$ DCs (Supplementary Information and Extended Data Fig. 7). Given that interactions both with DCs and with B cells are important for full Tfh cell differentiation[1,2], we suggest that T-cell EBI2 upregulation initially acts to favour interaction with Tfh-cell-promoting DCs and subsequently with activated B cells. While EBI2 upregulation is important for Tfh cell induction, the receptor is downregulated at week 2 of the response and this may facilitate Tfh cell retention in germinal centres[14]. Soluble CD25 was first detected in human serum ~30 years ago and it has since been reported in human and mouse serum in many studies and correlated with various disease conditions[21–23]. Although there has been evidence that sCD25 can antagonize certain IL-2 functions[10,22,24], the significance of sCD25 *in vivo* has been unclear. Moreover, the function of CD25 in myeloid cells has been mysterious[10,25,26], with some *in vitro* studies suggesting it suppresses[27] and others that it augments[28,29] T-cell responses. We show that DC production of CD25 plays an important role in quenching IL-2 in the outer T zone and it thereby cooperates with other factors, including ICOSL, to facilitate Tfh cell differentiation (Extended Data Fig. 7). While our findings show DCs produce sCD25, we do not exclude the possibility that membrane-associated CD25 on DCs also has a regulatory role. Strengths of IL-2 signalling influence Treg cell activity, Th17, Th1 and Th2 cell development, and CD8 T-cell proliferation and differentiation[10,30]. We suggest that CD25-mediated IL-2-quenching by DCs will be a general mechanism acting to guide a range of IL-2-sensitive cell activation and differentiation processes.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 2 February; accepted 29 March 2016.**

1. Ramiscal, R. R. & Vinuesa, C. G. T-cell subsets in the germinal center. *Immunol. Rev.* **252,** 146–155 (2013).
2. Crotty, S. T follicular helper cell differentiation, function, and roles in disease. *Immunity* **41,** 529–542 (2014).
3. Garside, P. *et al.* Visualization of specific B and T lymphocyte interactions in the lymph node. *Science* **281,** 96–99 (1998).
4. Haynes, N. M. *et al.* Role of CXCR5 and CCR7 in follicular Th cell positioning and appearance of a programmed cell death gene-1high germinal center-associated subpopulation. *J. Immunol.* **179,** 5099–5108 (2007).
5. Lee, S. K. *et al.* B cell priming for extrafollicular antibody responses requires Bcl-6 expression by T cells. *J. Exp. Med.* **208,** 1377–1388 (2011).
6. Johnston, R. J., Choi, Y. S., Diamond, J. A., Yang, J. A. & Crotty, S. STAT5 is a potent negative regulator of TFH cell differentiation. *J. Exp. Med.* **209,** 243–250 (2012).
7. Ballesteros-Tato, A. *et al.* Interleukin-2 inhibits germinal center formation by limiting T follicular helper cell differentiation. *Immunity* **36,** 847–856 (2012).
8. Nurieva, R. I. *et al.* STAT5 protein negatively regulates T follicular helper (Tfh) cell generation and function. *J. Biol. Chem.* **287,** 11234–11239 (2012).
9. Oestreich, K. J., Mohn, S. E. & Weinmann, A. S. Molecular mechanisms that control the expression and activity of Bcl-6 in TH1 cells to regulate flexibility with a TFH-like gene profile. *Nature Immunol.* **13,** 405–411 (2012).
10. Liao, W., Lin, J. X. & Leonard, W. J. Interleukin-2 at the crossroads of effector responses, tolerance, and immunotherapy. *Immunity* **38,** 13–25 (2013).

11. Pereira, J. P., Kelly, L. M., Xu, Y. & Cyster, J. G. EBI2 mediates B cell segregation between the outer and centre follicle. *Nature* **460,** 1122–1126 (2009).
12. Hannedouche, S. *et al.* Oxysterols direct immune cell migration via EBI2. *Nature* **475,** 524–527 (2011).
13. Liu, C. *et al.* Oxysterols direct B-cell migration through EBI2. *Nature* **475,** 519–523 (2011).
14. Suan, D. *et al.* T follicular helper cells have distinct modes of migration and molecular signatures in naive and memory immune responses. *Immunity* **42,** 704–718 (2015).
15. Meredith, M. M. *et al.* Expression of the zinc finger transcription factor zDC (Zbtb46, Btbd4) defines the classical dendritic cell lineage. *J. Exp. Med.* **209,** 1153–1165 (2012).
16. Goenka, R. *et al.* Cutting edge: dendritic cell-restricted antigen presentation initiates the follicular helper T cell program but cannot complete ultimate effector differentiation. *J. Immunol.* **187,** 1091–1095 (2011).
17. Yi, T. & Cyster, J. G. EBI2-mediated bridging channel positioning supports splenic dendritic cell homeostasis and particulate antigen capture. *eLife* **2,** e00757 (2013).
18. Yi, T. *et al.* Splenic dendritic cells survey red blood cells for missing self-CD47 to trigger adaptive immune responses. *Immunity* **43,** 764–775 (2015).
19. Logue, E. C., Bakkour, S., Murphy, M. M., Nolla, H. & Sha, W. C. ICOS-induced B7h shedding on B cells is inhibited by TLR7/8 and TLR9. *J. Immunol*. **177,** 2356–2364 (2006).
20. Marczynska, J. *et al.* The role of metalloproteinase ADAM17 in regulating ICOS ligand-mediated humoral immune responses. *J. Immunol.* **193,** 2753–2763 (2014).
21. Rubin, L. A. *et al.* Soluble interleukin 2 receptors are released from activated human lymphoid cells *in vitro*. *J. Immunol*. **135,** 3172–3177 (1985).
22. Maier, L. M. *et al.* Soluble IL-2RA levels in multiple sclerosis subjects and the effect of soluble IL-2RA on immune responses. *J. Immunol.* **182,** 1541–1547 (2009).
23. Chistiakov, D. A., Chistiakova, E. I., Voronova, N. V., Turakulov, R. I. & Savost'anov, K. V. A variant of the Il2ra/Cd25 gene predisposing to graves' disease is associated with increased levels of soluble interleukin-2 receptor. *Scand. J. Immunol.* **74,** 496–501 (2011).
24. Russell, S. E., Moore, A. C., Fallon, P. G. & Walsh, P. T. Soluble IL-2R$\alpha$ (sCD25) exacerbates autoimmunity and enhances the development of Th17 responses in mice. *PLoS ONE* **7,** e47748 (2012).
25. Kronin, V., Vremec, D. & Shortman, K. Does the IL-2 receptor $\alpha$ chain induced on dendritic cells have a biological function? *Int. Immunol.* **10,** 237–240 (1998).
26. Liang, D. *et al.* Role of CD25$^+$ dendritic cells in the generation of Th17 autoreactive T cells in autoimmune experimental uveitis. *J. Immunol.* **188,** 5785–5791 (2012).
27. Popov, A. *et al.* Infection of myeloid dendritic cells with *Listeria monocytogenes* leads to the suppression of T cell function by multiple inhibitory mechanisms. *J. Immunol.* **181,** 4976–4988 (2008).
28. Fukao, T. & Koyasu, S. Expression of functional IL-2 receptors on mature splenic dendritic cells. *Eur. J. Immunol.* **30,** 1453–1457 (2000).
29. Wuest, S. C. *et al.* A role for interleukin-2 trans-presentation in dendritic cell-mediated T cell activation in humans, as revealed by daclizumab therapy. *Nature Med.* **17,** 604–609 (2011).
30. Klatzmann, D. & Abbas, A. K. The promise of low-dose interleukin-2 therapy for autoimmune and inflammatory diseases. *Nature Rev. Immunol.* **15,** 283–294 (2015).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** We thank C. Allen, M. Ansel, T. Defranco, M. Muschen and S. Sanjabi for mice, J. An for help with the mouse colony, Y. Xu for help with quantitative PCR, and A. Abbas and M. Barnes for comments on the manuscript. J.G.C. is an investigator of the Howard Hughes Medical Institute. E.L. is supported by the University of California, San Francisco, Biomedical Sciences (BMS) Graduate program and the National Science Foundation (grant number 1144247). This work was supported in part by National Institutes of Health grant AI40098.

**Author Contributions** J.L. designed and performed experiments, interpreted the results and prepared the manuscript. E.L. performed several experiments including staining and quantitation of cell distribution in sections and helped prepare the manuscript. T.Y. performed experiments identifying the defects in EBI2 KO T cells. J.G.C. designed experiments, supervised research and wrote the manuscript.

**Author Information** Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.G.C. (jason.cyster@ucsf.edu).

## METHODS

**Mice and bone marrow chimaeras.** Wild-type C57BL/6NCr and C57BL/6-cBrd/cBrd/Cr (B6-Ly5.2) mice of 7–9 weeks of age were purchased from the National Cancer Institute. *Ebi2*$^{-/-}$ (containing a green fluorescent protein (GFP) reporter in place of the *Ebi2* coding exon), *Cyp7b1*$^{-/-}$, *Ch25h*$^{-/-}$, *Hsd3b7*$^{-/-}$, *Ccr7*$^{-/-}$, HEL-specific MD4 Ig-transgenic, HEL-specific Hy10 mice and OVA-specific OTII TCR-transgenic mice have been described (refs 11, 17, 31 and references therein). B-cell-deficient μMT mice were provided by T. Defranco and C. Allen. *Cd28*$^{-/-}$ mice were provided by K. M. Ansel. TCRβδ$^{-/-}$, *Cd47*$^{-/-}$, *Batf3*$^{-/-}$ and Zbtb46-DTR mice were from Jackson Laboratories. *Irf4*$^{fl/fl}$ CD11c-Cre$^{+}$ mice were from Jax and provided by S. Sanjabi. *Cd25*$^{-/-}$ mice were from Jax and provided by M. Muschen. BM chimaeras were generated as described[11] and analysed after 6–12 weeks. Mixed BM chimaeras were made by mixing equal amounts of the two types of BM before transfer. The sample sizes were guided by previous studies in our laboratory. No animals were excluded from analysis, and sample size estimates were not used. The mouse genotype was not blinded from the investigator. Mice of a given genotype were randomly assigned to groups. However, littermate mice were evenly distributed into control or treatment groups and mice of both groups were co-caged whenever possible. In experiments involving transfers of OTII T cells, since the TCR transgene is on the Y chromosome, male mice were used as donors and recipients. In other experiments, similar numbers of male and female mice were used. All mice were adult and were studied between 7 and 20 weeks of age. Animals were housed in a specific-pathogen free environment in the Laboratory Animal Research Center at the University of California, San Francisco, and all experiments conformed to ethical principles and guidelines approved by the Institutional Animal Care and Use Committee.

**Adoptive transfer, immunizations, DC ablation and treatments.** For analysis of CD4$^{+}$ T-cell position, activation or Tfh cell differentiation in spleens, $1 \times 10^6$ to $5 \times 10^6$ WT and/or EBI2 KO OTII cells were adoptively transferred into mice. One day after cell transfer, recipients were immunized intraperitoneally with $2 \times 10^8$ SRBCs (Colorado Serum Company) conjugated with OVA (Sigma-Aldrich) as described[32] with minor modifications detailed below, with 25 μg OVA plus 25 μg lipopolysaccharide (*Escherichia coli* 0111:B4, Sigma-Aldrich), intravenously with $2 \times 10^9$ heat-killed *Listeria*-OVA as described[33], or subcutaneously with 25 μg OVA in 200 μl Alum (InvivoGen). For conjugation of OVA with SRBCs, 1 ml of SRBCs was washed with PBS three times, incubated with 4 ml of 30 mg ml$^{-1}$ ice-cold OVA in PBS and crosslinked with 1 ml of 100 mg ml$^{-1}$ EDCI (1-ethyl-3-(3-dimethylaminopropyl) carbodiimide, Sigma-Aldrich) for 1 h on ice with occasional mixing, followed by washing four times in PBS to remove the free OVA and confirmation of the conjugation by flow cytometry. For HEL-specific antibody responses, $1 \times 10^5$ Hy10 B cells[34] were adoptively transferred into desired recipients. One day after cell transfer, recipients were intraperitoneally immunized with SRBCs conjugated with a low affinity mutant of HEL termed HEL2× (ref. 35) as described[17]. To visualize cell proliferation, cells were labelled with CellTrace violet tracer (Molecular Probes, Invitrogen) according to the manufacturer's instructions. For DC ablation, Zbtb46-DTR full or mixed chimaeras were injected intraperitoneally with 20 ng DT (Sigma-Aldrich) per gram of body weight 3 days before cell transfer and received 4 ng DT per gram of body weight on the third day after the initial DT injection and in some cases again 3 days later. To block ICOS, mice were injected intravenously with anti-mouse ICOS antibody (rat IgG2b, clone 7E.17G9, BioXCell) or rat IgG2b isotype control 1 day before and 2 days after cell transfer (0.5 mg per mouse per injection). To block IL-2, two doses of recombinant CD25 protein (25 μg per mouse per dose, Sino Biological) were injected into mice at the same time with immunization or 1 day after immunization.

**Flow cytometry and cell sorting.** All antibody conjugates were from Biolegend or BD Biosciences. EBI2 surface staining, T-cell staining, germinal centre B cell staining and intracellular Ig staining for plasma cells were performed as described[17,31]. EBI2 surface staining was performed with a goat polyclonal antibody against the amino (N) terminus (clone A20, Santa Cruz Biotechnology) as described[17]. Tfh cell staining was performed with antibodies including biotin-conjugated anti-CXCR5 (BD Biosciences), PE-Cy7-conjugated anti-PD-1 (clone RMP1-30, Biolegend) and Alexa 647-conjugated anti-Bcl6 (clone k112-91, BD Biosciences) as described[36]. Staining of CD25 on splenic or lymph node DC was performed using Alexa 647-conjugated anti-CD25 (clone PC61, Biolegend), PE-Cy7-conjugated anti-CD11c (clone N418, Tonbo Biosciences), FITC-conjugated anti-I-A$^{b}$ (clone AF6-120.1, BD Biosciences), PE-conjugated anti-DCIR2 (clone 33D1, eBioscience) or PE-conjugated anti-CD11b (clone M1/70, Biolegend), Pacific Blue-conjugated anti-CD8a (clone 53-6.7, Biolegend) and biotin-conjugated anti-CD103 (clone 2E7, Biolegend). To assess pSTAT5 levels directly *ex vivo*, spleens were immediately mashed using cell strainers into Cytofix/Cytoperm buffer (BD Biosciences). For peripheral lymph node analyses, the brachial, axillary and inguinal lymph node were pooled. After fixation for 30 min at 37 °C, the cells were washed, resuspended in Perm Buffer III (BD Biosciences) and incubated on ice for 30 min. After an additional wash, cells were stained for surface and intracellular antigens, including pSTAT5 (pY694, BD Biosciences), for 45 min at room temperature. Where indicated, 4 μg of human IL-2 was injected intravenously into mice 2 h before analysis as a positive control for pSTAT5. Data were collected on an LSRII and a FACSVerse (BD Biosciences) and were analysed with FlowJo software (TreeStar). DCs, OTII CD4$^{+}$ T cells or Tfh cells were sorted using a FACSAria III (BD) as described[31].

**Transwell migration assay.** Splenocytes were allowed to transmigrate for 4 h across 5 μm transwell filters (Corning Costar) towards medium or 7α,25-OHC (Avanti Polar Lipids) and enumerated by flow cytometry as described[17].

**Immunohistochemistry and immunofluorescence microscopy.** Cryosections of 7 μm were fixed and stained immunohistochemically as described[17,31] with: FITC-conjugated anti-IgD (clone 11-26c.2a, BD Biosciences), biotin-conjugated anti-CD45.1 (clone A20, Biolegend), biotin-conjugated anti-DCIR2 (clone 33D1, Biolegend) or biotin-conjugated anti-CD25 (clone PC61.5, BD Biosciences) followed by HRP-conjugated anti-FITC, AP-conjugated anti-FITC, and/or AP-conjugated SA (Jackson Immunoresearch). For staining of DCIR2 and CD25, a tyramide amplification kit was used (TSA Biotin System; Perkin Elmer). For immunofluorescence, staining was performed with biotin conjugated anti-CD45.1, rabbit anti-GFP (Molecular Probes), and goat anti-mouse IgD (GAM/IGD(FC)/7S, Cedarlane Laboratories), followed by AMCA-conjugated donkey anti-goat IgG (Jackson Immunoresearch), Alexa 488-conjugated donkey anti-rabbit IgG and Alexa 647-conjugated streptavidin (Invitrogen). Images were captured with a Zeiss AxioObserver Z1 inverted microscope.

**Image quantification.** Immunofluorescence images were analysed using IMARIS (version 7.3.0). White-pulp cords containing circular T zones were used to quantify outer T-zone positioning of co-transferred WT (red) and EBI2 HET or KO cells (green). OTII cells were defined using the Spots function in IMARIS and coordinates for each OTII cell were exported into R. The centre and average radius of the T zone was measured using the Measurement Points function in IMARIS. The 'outer T zone' was defined as the area further than three-quarters of the average radius from the centre of the T zone. The distance of each OTII cell from the centre of the T zone and the proportion of cells in the outer T zone was calculated using R. An average of 70 cells were present for each co-transferred group per T zone (average of 140 total). IHC images were analysed manually using the Cell Counter Plugin in ImageJ (version 1.49). OTII cells that were in contact with DCIR2$^{+}$ DC were distinguished from lone OTII cells using separate counters.

**Soluble CD25 ELISA and bioassay.** To test the production of sCD25 by DCs *in vitro*, mice were first intraperitoneally immunized with SRBCs. At the time of analysis, spleen CD4$^{+}$ DCs were enriched by depletion of T, B and natural killer (NK) cells with a cocktail of biotin-conjugated antibodies and isolated by positive selection using biotin-conjugated anti-DCIR2 to purities of over 90% (Miltenyi Biotec). The purified DCs were cultured *in vitro* for 8 h and the presence of soluble CD25 in culture supernatants was detected using a CD25 ELISA kit (R&D Systems). To detect sCD25 in spleen tissue, each spleen was mashed into 1 ml of medium through a 70 μm cell strainer and centrifuged at 300*g* for 10 min and 3,000*g* for 15 min at 4 °C. The cell-free supernatant was subjected to the ELISA assay. To test for antagonism of IL-2 mediated signalling, DC culture supernatants were mixed with different dosages of recombinant mouse IL-2 (Biolegend) for 2 h and added to splenocytes at 37 °C for 30 min. pSTAT5 levels in CD25$^{+}$CD4$^{+}$ T cells were analysed as described above.

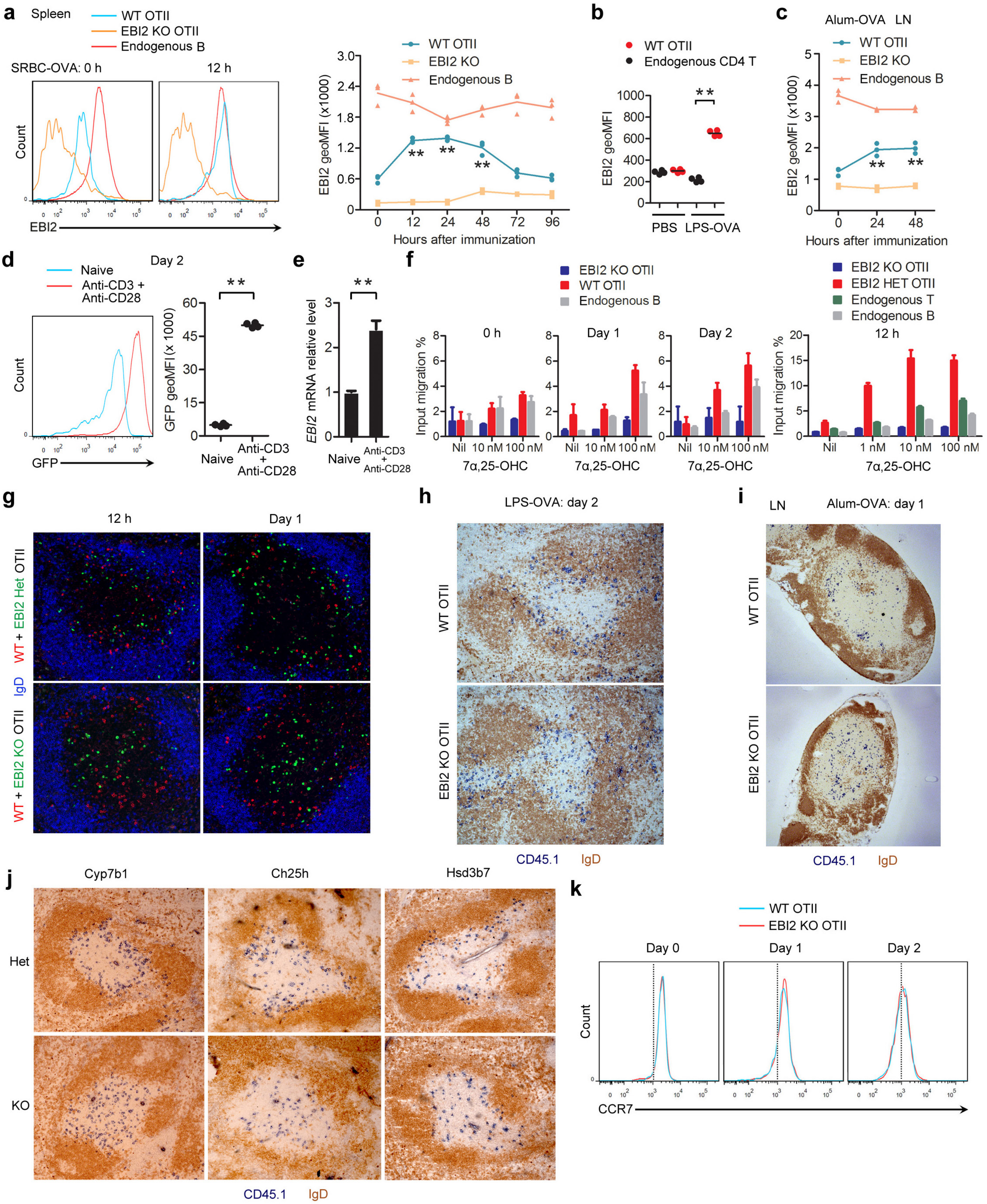
**Detection of SRBC- or HEL-specific antibody responses.** To assay for anti-SRBC or anti-HEL IgM and IgG from mouse serum, 50 μl of SRBCs or HEL-conjugated mouse RBCs ($5 \times 10^7$ cells per millilitre in PBS) were incubated with 2 μl of serum for 1 h at room temperature. After washing, the RBCs were incubated with fluorescent antibodies against mouse IgM, IgG1 and IgG2b for flow cytometric analysis.

**Quantitative RT–PCR.** Total RNA from sorted cells was isolated and reverse-transcribed, and quantitative PCR was performed as described[31]. Data were analysed using the comparative $C_T$ ($2^{-\Delta\Delta C_T}$) method using *Hprt* as the reference.

**RNA-seq analysis.** Spleens were taken 1 h after saline or SRBC immunization. CD4$^{+}$ DCs were pre-enriched using MACS manual cell separation columns with anti-CD11c microbeads (Miltenyi Biotec) and further sorted on the basis of surface markers of CD11c$^{+}$I-Ab$^{+}$CD4$^{+}$CD8$^{-}$. Cells were sorted twice on a FACSAira III to purities of over 99%. Sorted DCs ($10^6$) were snap frozen and then RNA was extracted with the QIAGEN RNeasy Kit. RNA quality was checked with an Agilent 2100 Bioanalyzer (RNA integrity number >9 for all samples). Barcoded sequencing libraries were generated with 100 ng of RNA with an Ovation RNA-Seq System V2 and Encore Rapid Library System. Sequencing was performed on an Illumina HiSeq 2500 (UCSF Human Genetics Core) with 100-base-pair paired-end reads. Sequences were reported as FASTQ files, which were aligned to the mm9 mouse genome with STAR (Spliced Transcript Alignment to a Reference). Generation of $\log_2$FC values and further analyses were performed with a Bioconductor package on RStudio. The RNA-seq data have been deposited in the Gene Expression Omnibus (NCBI) data repository under accession code GEO: GSE71165.

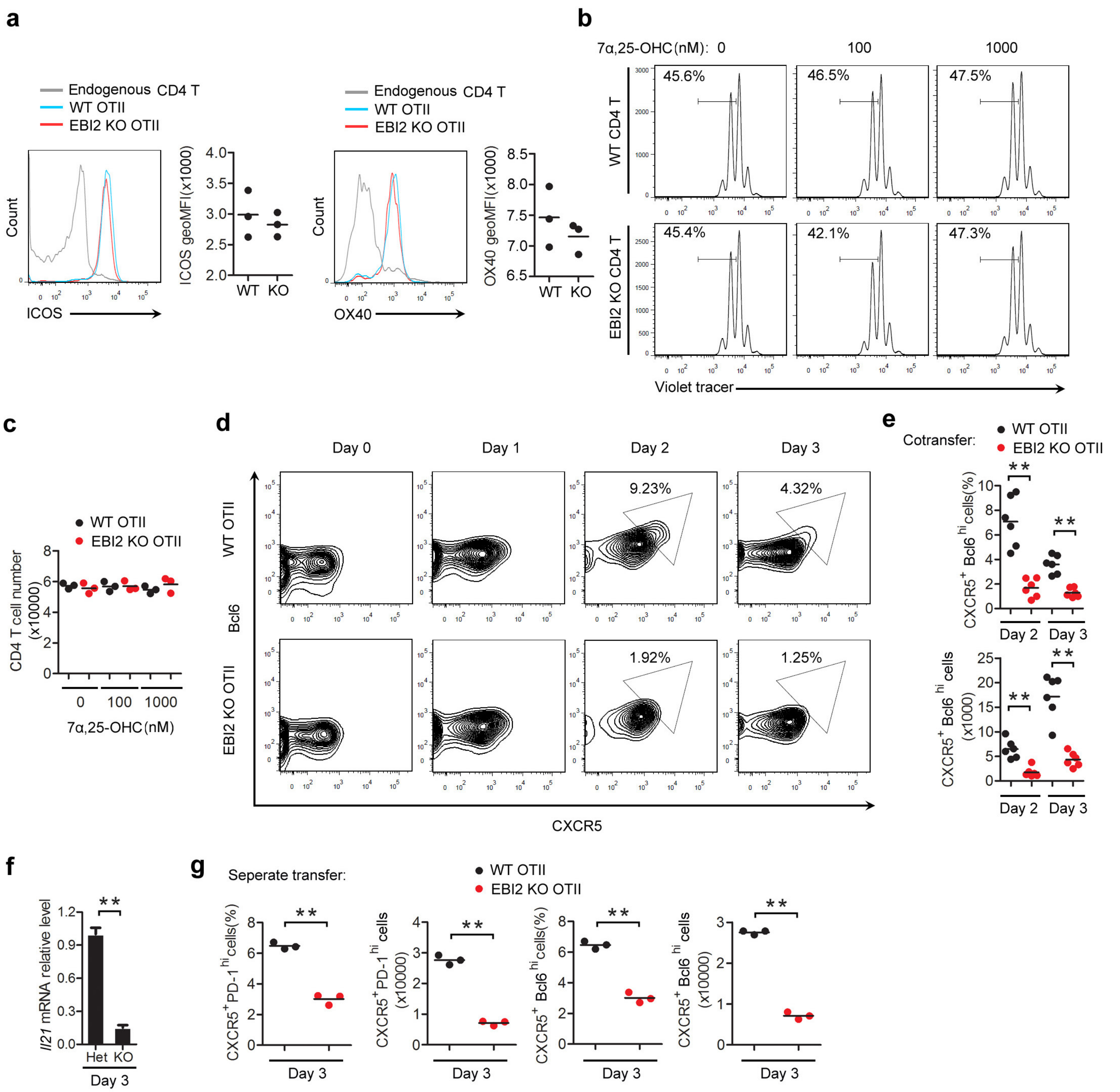
**Statistical analysis.** No statistical methods were used to predetermine sample size. Prism software (GraphPad) was used for all statistical analysis. Statistical comparisons were performed using an ANOVA or a two-tailed Student's *t*-test. *P* values were considered significant when less than 0.05. In all summary dot plots, points indicate data from individual mice, and horizontal lines indicate means. In bar graphs, bars indicate means, and error bars indicate s.e.m.


31. Yi, T. *et al.* Oxysterol gradient generation by lymphoid stromal cells guides activated B cell movement during humoral responses. *Immunity* **37,** 535–548 (2012).
32. Carlsson, F., Getahun, A., Rutemark, C. & Heyman, B. Impaired antibody responses but normal proliferation of specific $CD4^+$ T cells in mice lacking complement receptors 1 and 2. *Scand. J. Immunol.* **70,** 77–84 (2009).
33. Muraille, E. *et al.* Distinct in vivo dendritic cell activation by live versus killed *Listeria monocytogenes*. *Eur. J. Immunol.* **35,** 1463–1471 (2005).
34. Allen, C. D., Okada, T., Tang, H. L. & Cyster, J. G. Imaging of germinal center selection events during affinity maturation. *Science* **315,** 528–531 (2007).
35. Paus, D. *et al.* Antigen recognition strength regulates the choice between extrafollicular plasma cell and germinal center B cell differentiation. *J. Exp. Med.* **203,** 1081–1091 (2006).
36. Meli, A. P. & King, I. L. Identification of mouse T follicular helper cells by flow cytometry. *Methods Mol. Biol.* **1291,** 3–11 (2015).

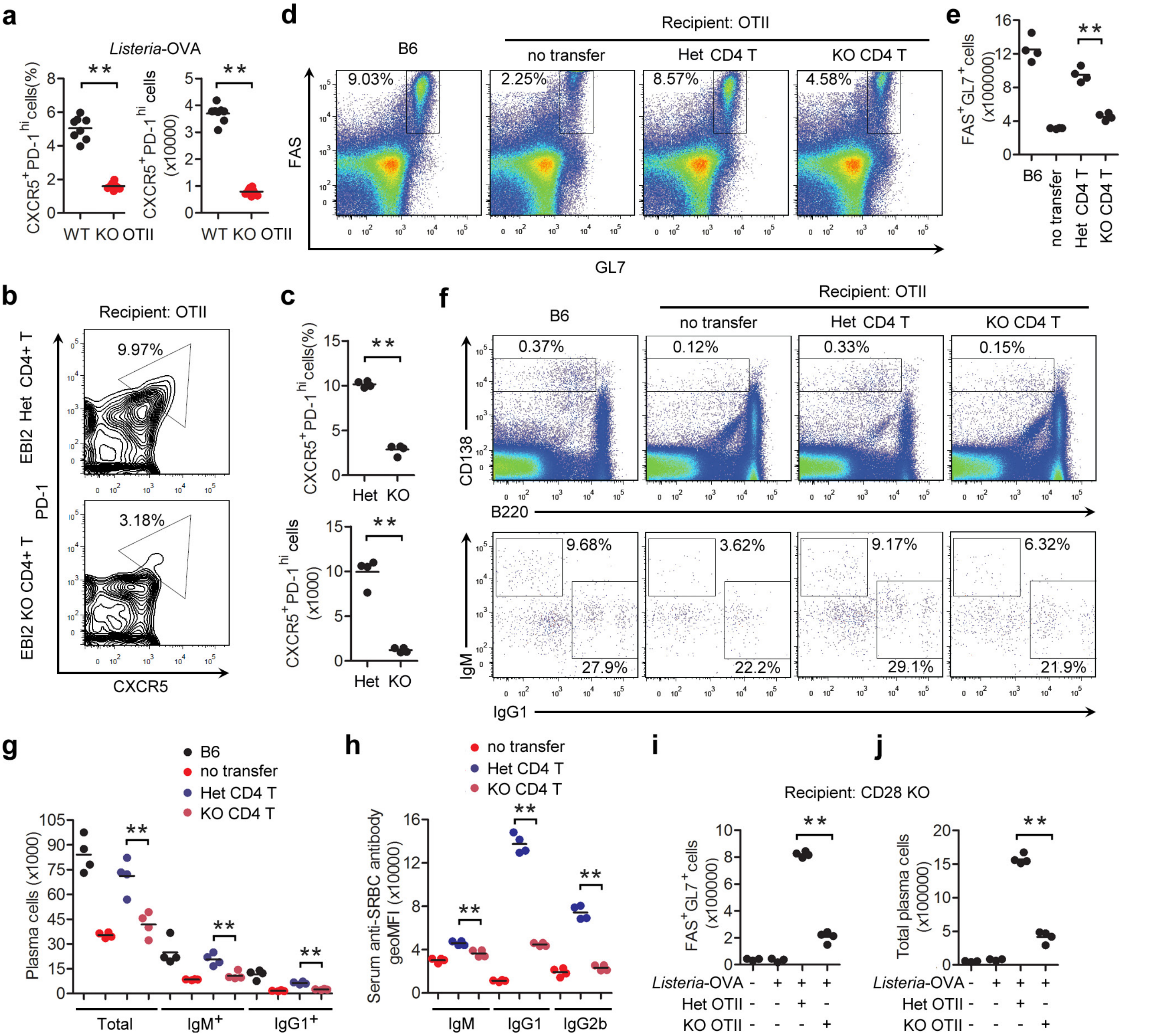



**Extended Data Figure 1** | See next page for caption.

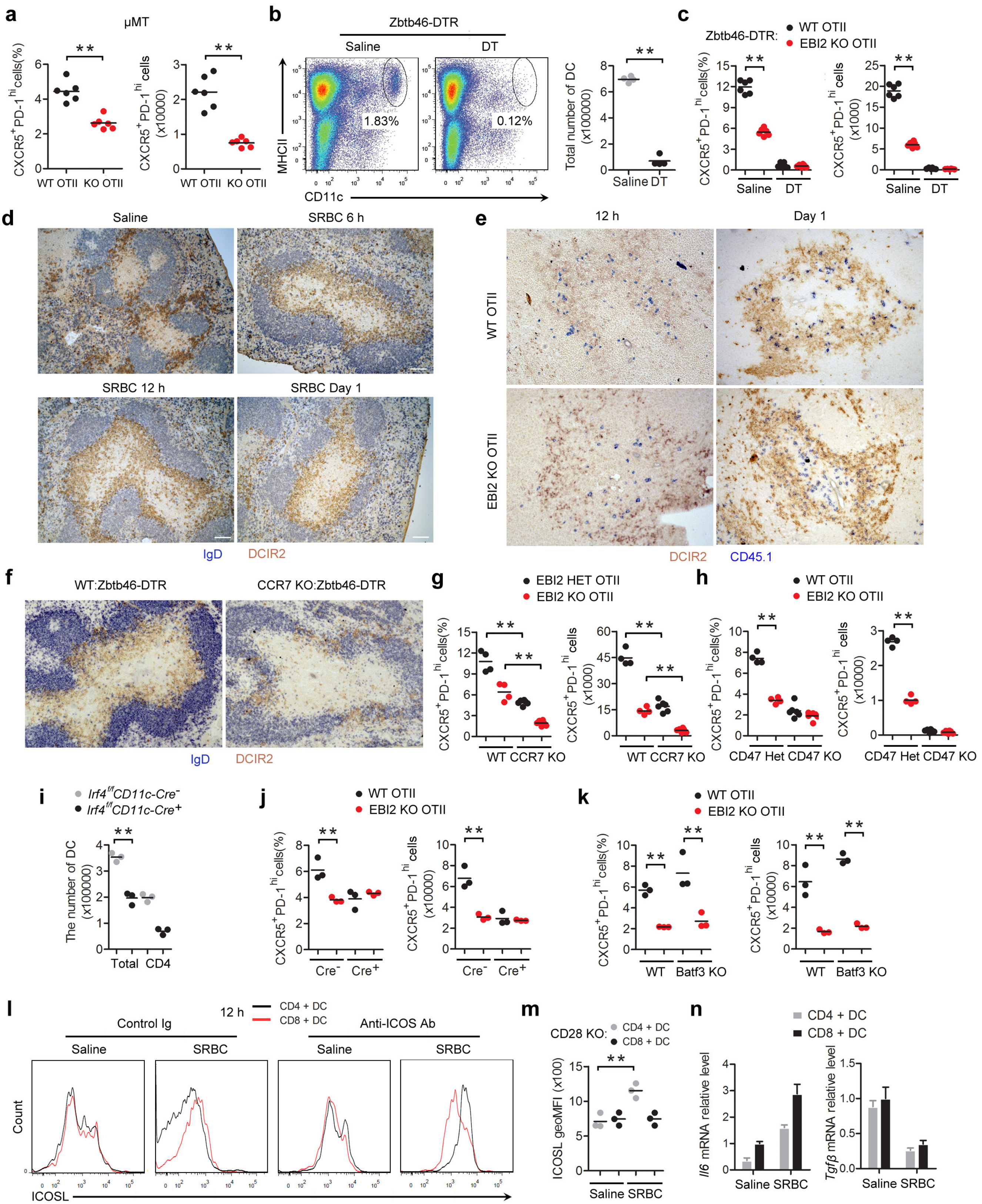
**Extended Data Figure 1 | EBI2 and 7α,25-OHC promote positioning of newly activated CD4 T cells in the outer T zone. a**, Flow cytometric analysis of EBI2 expression on splenic OTII T cells and endogenous B cells in transfer recipients at 0 and 12 h after SRBC-OVA immunization. EBI2 KO cells were used as a staining control. Left histograms show example FACS data and right panel shows summary data across the indicated time points as geometric mean fluorescence intensity (geoMFI). **b**, EBI2 expression on OTII and endogenous T cells in transfer recipients 2 days after saline or lipopolysaccharide-OVA immunization. Left histograms show example flow cytometric data and right panel shows summary geoMFI data for four mice. **c**, Summary geoMFI time course data of EBI2 expression on lymph node OTII T cells in transfer recipients at the indicated times after alum-OVA immunization. **d**, GFP expression in EBI2$^{GFP/+}$ CD4 T cells that were unstimulated (naive) or treated with anti-CD3 plus anti-CD28 for 2 days. Left histogram shows example flow cytometric data and right panel shows summary geoMFI data for three mice. **e**, *Ebi2* mRNA abundance in cells of the type in **d**, determined by RT–qPCR and shown relative to the naive cells. **f**, Migration of OTII T cells and endogenous cells to the indicated amounts of 7α,25-OHC in transwell assays. Cells were from unimmunized (0 h) or immunized (day 1, 2) transfer recipient mice in one experiment (left) or from 12 h immunized transfer recipients in a second experiment (right). Data are shown as percentage of input cells of each type that migrated. **g**, Immunofluorescence analysis of spleen showing the distribution of co-transferred WT CD45.1$^+$ (red) and EBI2 het or KO (GFP$^+$, green) OTII T cells and endogenous B cells (IgD, blue) at 12 h and 1 day after immunization. **h**, **i**, Immunohistochemical analysis of WT spleens (**h**) and inguinal lymph nodes (**i**) showing the distribution of transferred control (WT) or EBI2 deficient (KO) OTII CD45.1$^+$ T cells (blue) and endogenous B cells (IgD, brown) at day 2 after lipopolysaccharide-OVA immunization (**h**) or day 1 after alum-OVA immunization (**i**). **j**, Immunohistochemical analysis of Cyp7b1, Ch25h or Hsd3b7 control (het, upper panels), or KO (lower panels) spleens showing the distribution of transferred WT OTII T cells (CD45.1, blue) and endogenous B cells (IgD, brown) at day 2 after SRBC-OVA immunization. **k**, CCR7 expression on WT and EBI2 KO OTII T cells in transfer recipient spleens at the indicated days after SRBC-OVA immunization. $^{**}P < 0.01$ by ANOVA (**a**, **c**) or Student's *t*-test (**b**, **d**, **f**). Data are representative of two (**a**–**i**, **k**) or three (**j**) independent experiments with at least three (**a**–**c**, **k**) or two (**d**–**j**) mice per group (error bars (**e**), s.e.m.).




**Extended Data Figure 2 | Defective differentiation of EBI2-deficient T cells to follicular helpers.** **a**, ICOS and OX40 expression on WT, EBI2-deficient (KO) OTII and endogenous CD4 T cells in transfer recipient spleens two days after SRBC-OVA immunization. **b**, **c**, *In vitro* proliferation of WT and EBI2 KO T cells in response to anti-CD3 plus anti-CD28 in the presence of the indicated amounts of 7α,25-OHC, shown as violet tracer dye dilution profiles (**b**) and total CD4 T-cell numbers (**c**) at day 3 of culture. Numbers in **b** indicate frequency of cells that have undergone two or more divisions. **d**, Flow cytometric analysis of co-transferred WT and EBI2 KO OTII T cells for CXCR5 and intracellular Bcl6 expression at the indicated days after SRBC-OVA immunization. Numbers indicate frequency of cells in gated region. **e**, Summary of data of the type in **d**. Upper plot shows frequency and lower plot number of $CXCR5^{+}Bcl6^{hi}$ OTII T cells. **f**, *Il21* mRNA abundance in $CXCR5^{+}$ $PD\text{-}1^{hi}$ control (Het) or EBI2 KO OTII T cells sorted from recipient spleens at day 3 after immunization with SRBC-OVA, determined by RT–qPCR and shown relative to the Het control. **g**, Frequency and number of $CXCR5^{+}$ $PD\text{-}1^{hi}$ (left) or $CXCR5^{+}$ $Bcl6^{hi}$ (right) WT and EBI2 KO OTII T cells in mice that received the cells as separate transfers, at day 3 after SRBC-OVA immunization. $^{**}P < 0.01$ by ANOVA (**e**) or Student's *t*-test (**f**, **g**). Data are representative of three (**a**, **d**, **e**, **g**) or two (**b**, **c**, **f**) independent experiments with at least three mice per group (error bars (**f**), s.e.m.).

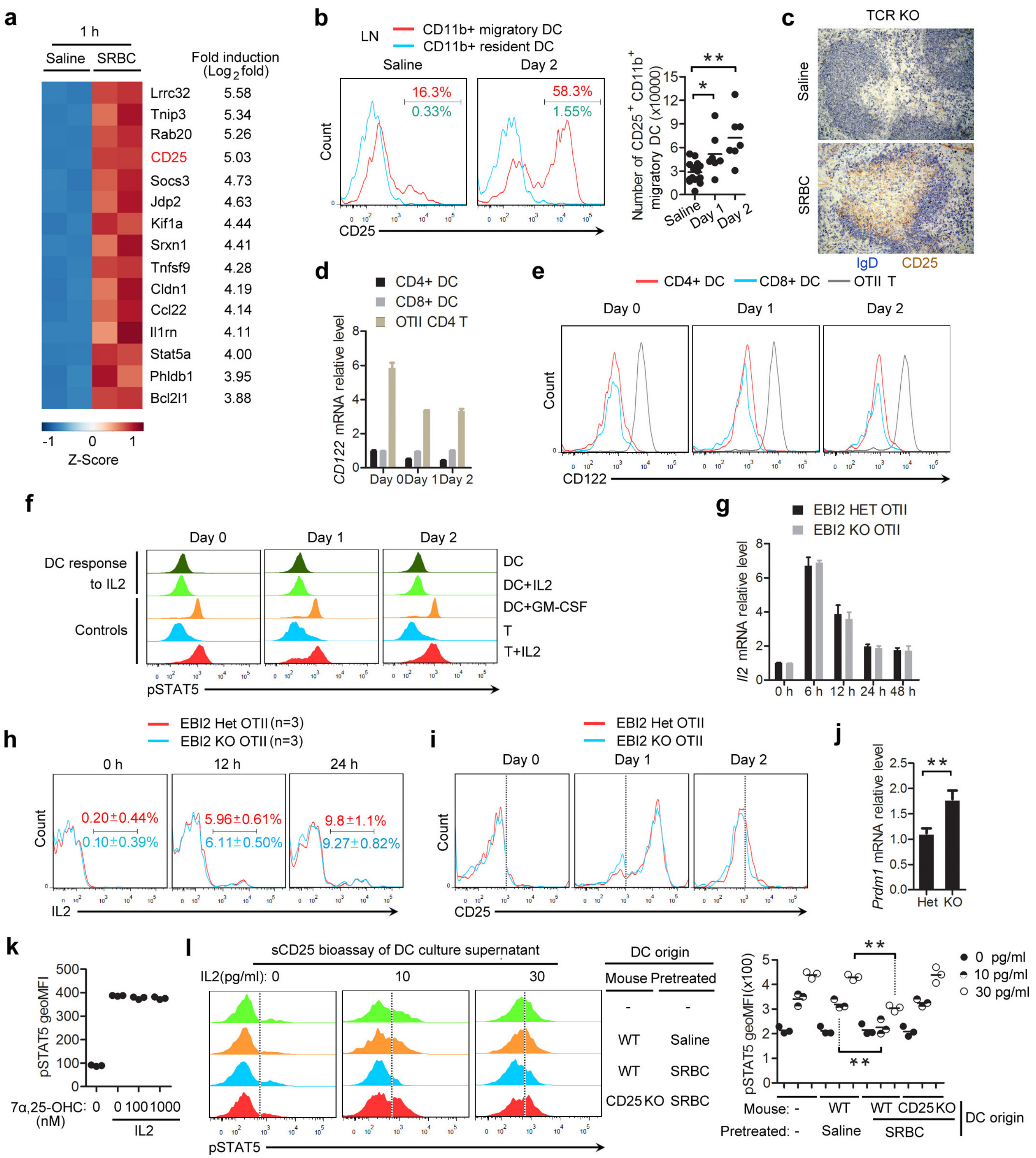



**Extended Data Figure 3 | EBI2-deficient T cells support reduced plasma cell and germinal centre response. a**, Frequency and number of CXCR5$^+$PD-1$^{hi}$ CD4$^+$ control (WT) and EBI2 KO OTII T cells in spleens of day 3 *Listeria*-OVA immunized transfer recipients. **b**, PD-1 and CXCR5 flow cytometric analysis of control (Het) and EBI2 KO polyclonal CD4 T cells co-transferred to OTII recipients, 8 days after immunization with unconjugated SRBCs. **c**, Summary of data of the type in **b** shown as CXCR5$^+$PD-1$^{hi}$ cell frequency and number. **d**, Flow cytometric analysis for the germinal centre markers FAS and GL7 on endogenous B cells in OTII TCR transgenic mice that received no cells, control (HET) CD4 T cells or EBI2 KO CD4 T cells, or in WT B6 control mice, 12 days after immunization with unconjugated SRBCs. **e**, Summary of data from **d** shown as number of FAS$^+$GL7$^+$ germinal centre B cells. **f**, Flow cytometric analysis for CD138$^{hi}$ B220$^{lo}$ plasma cells (top) and intracellular IgM and IgG1 staining of these cells (bottom) in mice of the type in **d**. **g**, Summary of data from **f** shown as number of cells. **h**, Serum anti-SRBC antibody levels in mice of the type in **d**, determined by flow cytometric analysis of SRBCs stained with immune sera, plotted as geoMFI. **i**, **j**, Number of Fas$^+$GL7$^+$ germinal centre cells and CD138$^{hi}$B220$^{int}$ plasma cells in *Listeria*-OVA immunized CD28 KO mice that had received control (het) or EBI2 KO OTII cells, analysed at day 5. $^{**}P < 0.01$ by ANOVA (**g**, **h**) or Student's *t*-test (**a**, **e**, **c**, **i**, **j**). Data are representative of two independent experiments with at least three mice per group.



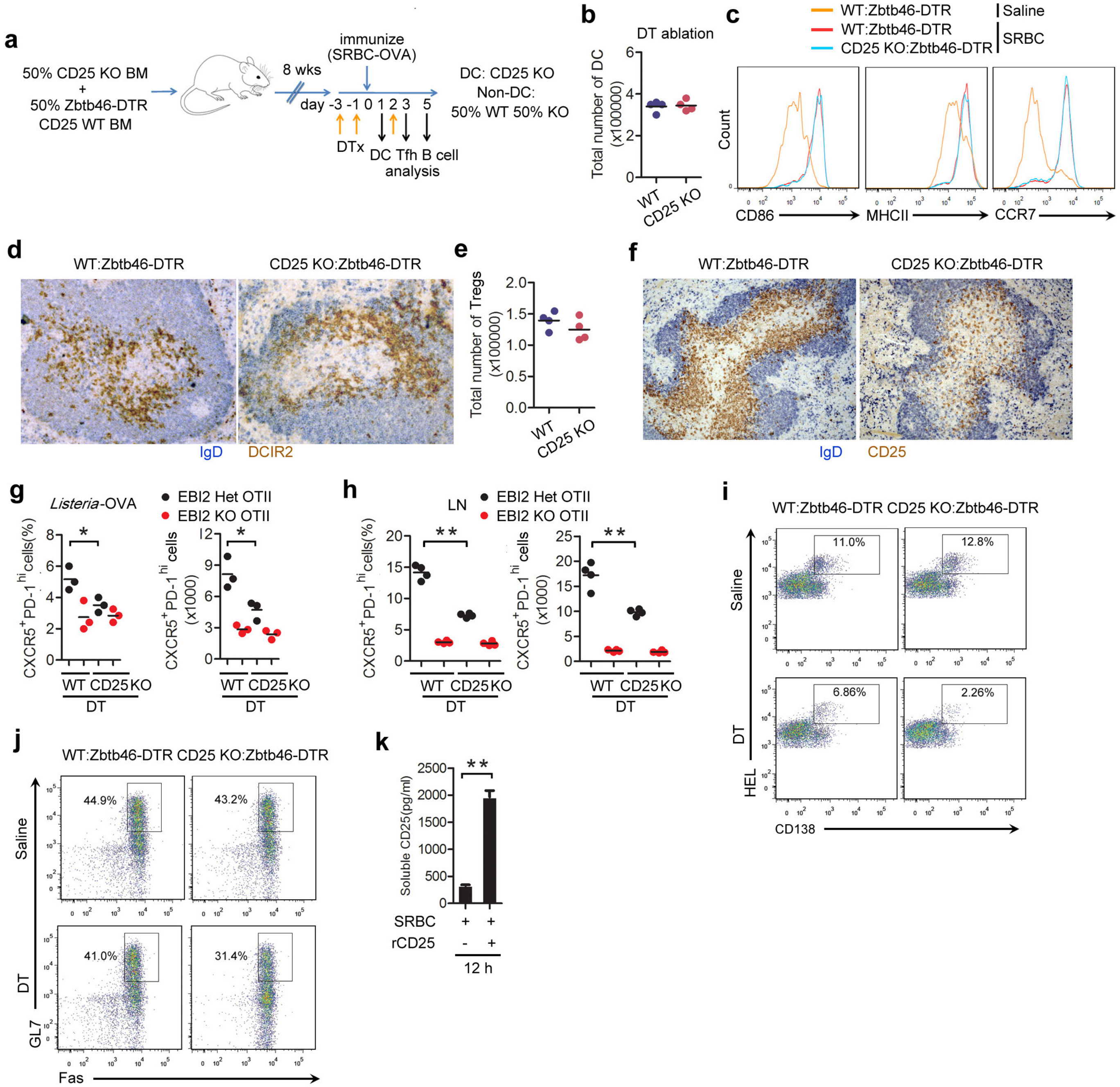
**Extended Data Figure 4 |** See next page for caption.

**Extended Data Figure 4 | T-cell EBI2 is required for CD4$^+$ DC-mediated augmentation of Tfh cell induction.** **a**, Frequency and number of CXCR5$^+$PD-1$^{hi}$ OTII T cells in μMT recipients determined by flow cytometric analysis. **b**, Flow cytometric analysis for CD11c and MHC class II on splenocytes from Zbtb46-DTR mice treated with saline or DT for 1 day. Graph shows summary data for DC number in four mice of each type. **c**, Frequency and number CXCR5$^+$PD-1$^{hi}$ WT and EBI2 KO OTII T cells in spleens from Zbtb46-DTR BM chimaeras treated with saline or diphtheria toxin (DT), at day 3 after immunization with SRBC-OVA. **d**, Immunohistochemical analysis of spleen sections from WT mice without immunization (saline) or SRBC immunized for the indicated times, stained to detect IgD$^+$ B cells (blue) and DCIR2$^+$ DCs (brown). **e**, Immunohistochemical analysis of spleen sections from recipients of WT or EBI2 KO OTII T cells at 12 h and 1 day after immunization SRBC-OVA immunization, stained for OTII CD45.1$^+$ T cells (blue) and DCIR2$^+$ DCs (brown). **f**, Immunohistochemical analysis of spleen sections from WT:Zbtb46-DTR or CCR7 KO:Zbtb46-DTR mixed BM chimaeras treated with DT, at day 2 after immunization. **g**, Frequency and number of CXCR5$^+$PD-1$^{hi}$ WT and EBI2 KO OTII T cells in spleens from WT:Zbtb46-DTR (control) or CCR7 KO:Zbtb46-DTR mixed BM chimaeras treated with DT, at day 3 after immunization. **h**, Frequency and number of CXCR5$^+$PD-1$^{hi}$ control (het) and EBI2 KO co-transferred OTII T cells in spleens of CD47 KO recipients at day 3 after SRBC-OVA immunization. **i**, Number of total and CD4$^+$ DC in spleens from Irf4$^{f/f}$ CD11c-Cre$^-$ or $^+$ mice. **j**, As for **g** but in Irf4$^{f/f}$ CD11c-Cre$^-$ or $^+$ recipient mice. **k**, As for **g** but in Batf3 KO recipient mice. **l**, ICOSL surface levels for DCs from mice immunized 12 h earlier with saline or SRBCs and treated with control or ICOS blocking antibody. **m**, ICOSL surface levels for CD4$^+$ or CD8$^+$ DCs from CD28 KO mice immunized 12 h earlier with saline or SRBCs. **n**, *Il6* and *Tgfβ* mRNA abundance in sorted CD4$^+$ and CD8$^+$ splenic DCs from mice treated with saline or SRBC 6 h earlier, determined by RT–qPCR, shown relative to the control CD8$^+$ DC. $^{**}P < 0.01$ by ANOVA (**g**, **k**) or Student's *t*-test (**a**–**c**, **i**, **j**, **m**). Data are representative of three (**a**–**e**) or two (**f**–**n**) independent experiments with at least three (**a**–**c**, **g**–**n**) or two (**d**–**f**) mice per group (error bars (**n**), s.e.m.).

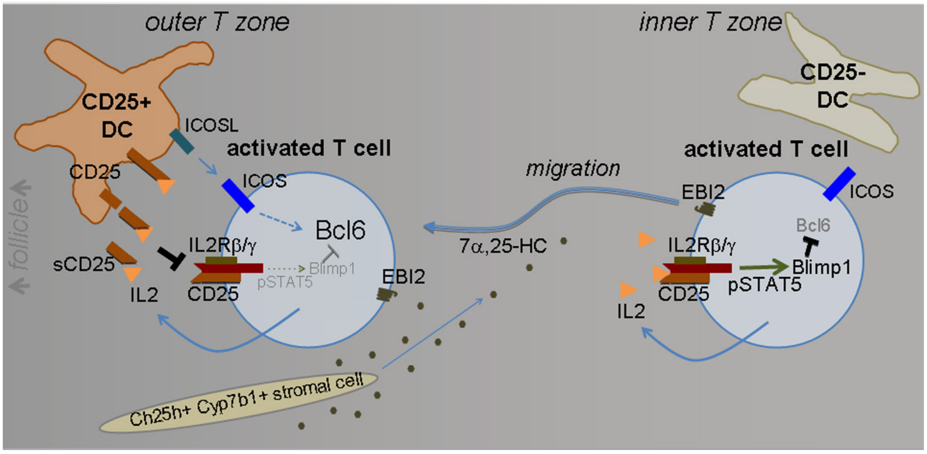



**Extended Data Figure 5** | See next page for caption.

**Extended Data Figure 5 | DCs produce membrane and soluble CD25 and inhibit IL-2R signalling.** **a**, Heat map of RNA-seq data from sorted $CD4^+$ splenic DC showing the top 15 most induced genes at 1 h after SRBC versus saline immunization. **b**, CD25 surface levels in $CD11b^+$ migratory and resident DCs from lymph nodes of mice immunized with saline or alum-OVA 2 days earlier. Graph on right shows summary data for total number of migratory $CD25^+$ DCs. **c**, Immunohistochemical analysis of spleen sections from TCRβδ KO mice immunized 1 day earlier with saline or SRBC, stained to detect IgD (blue) and CD25 (brown). **d**, **e**, CD122 (Il2rb) mRNA determined by RT–qPCR (**d**) and surface staining (**e**) on the indicated cell types isolated from spleens of WT OTII T-cell recipients at day 0, 1 and 2 after SRBC-OVA immunization. Transcript data are plotted relative to the signal in $CD4^+$ DCs at day 0. **f**, Intracellular flow cytometric analysis of pSTAT5 in $CD4^+$ DCs or, as a positive control $CD4^+$ T cells, that were untreated or incubated with IL-2 (200 pg $ml^{-1}$) or, as a further positive control, GM-CSF (100 pg $ml^{-1}$). **g**, *Il2* mRNA in control (Het) and EBI2 KO OTII T cells isolated from recipient mice at the indicated times after SRBC-OVA immunization. **h**, Intracellular flow cytometry for IL-2 in cells of the type in **f** at 0, 12 and 24 h. Percentages show mean (±s.e.m.) for three mice at each time point. **i**, Flow cytometric analysis of CD25 expression on co-transferred WT and EBI2 KO OTII T cells in WT recipients at the indicated days after SRBC-OVA immunization. **j**, *Prdm1* (encoding Blimp1) transcript levels in sorted $CXCR5^+PD\text{-}1^{hi}$ control (het) and EBI2 KO OTII T cells from SRBC-OVA immunized mice at day 3, plotted relative to the mean level in the Het group. **k**, Summary of pSTAT5 staining data for OTII T cells from mice immunized 1 day earlier with SRBC-OVA, incubated with the indicated amounts of 7α,25-OHC plus IL-2 (200 pg $ml^{-1}$) for 1 h. **l**, Flow cytometry of pSTAT5 in $CD25^+$ (regulatory) T cells exposed to the indicated amounts of IL-2 that had been pre-mixed with supernatants (s/n) from 8 h cultures of splenic $CD4^+$ DCs from WT or CD25 KO mice immunized with saline or SRBCs 1 day before. Graph on right shows summary data from one experiment. $^{**}P < 0.01$ by ANOVA (**b**, **l**) or Student's *t*-test (**j**). Data are representative of one (**a**) or two (**b**–**l**) independent experiments with at least two (**a**) or three (**b**–**l**) mice per group (error bars (**g**, **j**), s.e.m.).




**Extended Data Figure 6 | DC CD25 expression reduces IL-2 signalling in activated CD4 T cells, favouring their differentiation to follicular helpers. a**, Diagram of CD25 KO:Zbtb46-DTR BM chimaera generation and time line of experiment. DTx, DT treatment. **b**–**d**, Numbers (**b**), surface marker expression (**c**) and outer T zone positioning (**d**) of CD4$^+$ DCIR2$^+$ DCs in WT:Zbtb46-DTR and CD25 KO:Zbtb46-DTR mixed BM chimaeras pre-treated with DT, at day 1 after saline or SRBC immunization. **e**, Number of Foxp3$^+$ CD25$^+$ regulatory T cells in mice of the type in **b** except that the mice were immunized for three days. **f**, Immunohistochemical analysis of spleen sections from mice of the type in **b**, stained to detect IgD (blue) and CD25 (brown). **g**, **h**, Frequency and number of CXCR5$^+$PD-1$^{hi}$ control (EBI2 Het) and EBI2 KO OTII T cells in spleens (**g**) or lymph nodes (**h**) from WT:Zbtb46-DTR or CD25 KO:Zbtb46-DTR mixed BM chimaeras pre-treated with saline or DT, at day 3 after immunization with *Listeria*-OVA (**g**) or alum-OVA (**h**). **i**, **j**, Flow cytometric analysis for HEL-binding CD138$^+$ plasma cells (**i**) and HEL-binding GL7$^+$ Fas$^+$ germinal centre B cells (**j**) in spleens from WT:Zbtb46-DTR (control) or CD25 KO:Zbtb46-DTR mixed BM chimaeras that had received Hy10 B cells and been treated with DT, at day 5 after immunization with HEL-SRBC. **k**, Soluble CD25 detected by ELISA in spleen extracts taken from 12 h SRBC immunized mice, at day 1 after saline or recombinant CD25 treatment. $^{**}P < 0.01$ by Student's *t*-test (**h**, **k**). Data are representative of two independent experiments with at least three (**b**, **c**, **g**–**k**) or two (**d**, **f**) mice per group (error bars (**k**), s.e.m.).




**Extended Data Figure 7 | Model of how EBI2-dependent positioning of activated T cells in association with CD25$^{+}$ DCs in the outer T zone favours Tfh cell differentiation.** Initially, cognate T cells throughout the T zone are activated by antigen recognition and promptly start upregulating EBI2 and making IL-2. EBI2 guides cells to the 7α,25-OHC high outer T zone and in this location they interact with activated DCs producing membrane and shed CD25 that binds and quenches IL-2. This limits IL-2R signalling on the T cell via pSTAT5 and allows induction of Bcl6 by other inputs such as ICOSL. T cells that lack EBI2 or remain in the inner T zone for other reasons are exposed to autocrine IL-2 and this induces Blimp1, a repressor of Bcl6 (ref. 2), disfavouring the Tfh cell fate.

doi:10.1038/nature17950

# Noncanonical autophagy inhibits the autoinflammatory, lupus-like response to dying cells

Jennifer Martinez[1,2], Larissa D. Cunha[1], Sunmin Park[3], Mao Yang[1], Qun Lu[3], Robert Orchard[3], Quan-Zhen Li[4], Mei Yan[4], Laura Janke[1], Cliff Guy[1], Andreas Linkermann[5], Herbert W. Virgin[3] & Douglas R. Green[1]

**Defects in clearance of dying cells have been proposed to underlie the pathogenesis of systemic lupus erythematosus (SLE)[1]. Mice lacking molecules associated with dying cell clearance develop SLE-like disease[2], and phagocytes from patients with SLE often display defective clearance and increased inflammatory cytokine production when exposed to dying cells *in vitro*. Previously, we[3–6] and others[7] described a form of noncanonical autophagy known as LC3-associated phagocytosis (LAP), in which phagosomes containing engulfed particles, including dying cells[3,4,7], recruit elements of the autophagy pathway to facilitate maturation of phagosomes and digestion of their contents. Genome-wide association studies have identified polymorphisms in the *Atg5* (ref. 8) and possibly *Atg7* (ref. 9) genes, involved in both canonical autophagy and LAP[3–7], as markers of a predisposition for SLE. Here we describe the consequences of defective LAP *in vivo*. Mice lacking any of several components of the LAP pathway show increased serum levels of inflammatory cytokines and autoantibodies, glomerular immune complex deposition, and evidence of kidney damage. When dying cells are injected into LAP-deficient mice, they are engulfed but not efficiently degraded and trigger acute elevation of pro-inflammatory cytokines but not anti-inflammatory interleukin (IL)-10. Repeated injection of dying cells into LAP-deficient, but not LAP-sufficient, mice accelerated the development of SLE-like disease, including increased serum levels of autoantibodies. By contrast, mice deficient in genes required for canonical autophagy but not LAP do not display defective dying cell clearance, inflammatory cytokine production, or SLE-like disease, and, like wild-type mice, produce IL-10 in response to dying cells. Therefore, defects in LAP, rather than canonical autophagy, can cause SLE-like phenomena, and may contribute to the pathogenesis of SLE.**

LAP is a process in which some, but not all components of the autophagy machinery conjugate LC3 to phosphatidylethanolamine directly on the phagosome membrane[3,4,6,7,10]. The lipidated LC3 (LC3-II) then facilitates lysosomal fusion and cargo destruction. Both LAP and canonical autophagy require ATG7, ATG3, ATG5, ATG12, and ATG16L for LC3 lipidation[4,5]. However, unlike autophagy, LAP proceeds independently of the pre-initiation complex containing ULK1 and FIP200 (also known as RB1CC1)[3,4,6,7], and uses a distinct beclin 1 (BECN1) and VPS34 complex complex lacking ATG14 (ref. 5). By contrast, LAP, but not canonical autophagy, requires NADPH oxidase-2 (NOX2)[5], and rubicon (RUBCN)[5]. These requirements for LAP and canonical autophagy can therefore distinguish between the two processes (Supplementary Table 1). As many components of autophagy are required for development (for example, FIP200 (refs 11, 12) and BECN1 (ref. 12)) or post-natal survival (for example, ATG14 (refs 12, 13), ATG7 (ref. 12), ATG5 (ref. 12) and ATG16L (ref. 12)), we generated animals in which several autophagy genes were conditionally ablated using lysozyme M (LysM, also known as Lyz2)-Cre recombinase[14], affecting macrophages (CD11b$^+$ F4/80$^+$), monocytes (CD11b$^+$ CD115$^+$), some neutrophils (CD11b$^+$ Ly6G$^+$), and some conventional dendritic cells (CD11b$^+$ CD11c$^+$), but not eosinophils, plasmacytoid dendritic cells, or lymphocytes (Extended Data Fig. 1a, b). While all animals appeared normal at weaning, we observed that LAP-deficient genotypes failed to gain weight compared to their wild-type littermates (Fig. 1a). This effect was observed in animals lacking proteins required for both LAP and autophagy (ATG7, ATG5, BECN1) or LAP alone (NOX2, RUBCN), but not in animals lacking proteins required for autophagy but dispensable for LAP (FIP200, ULK1). Compared to LAP-sufficient animals, LAP-deficient mice displayed increased levels of circulating lymphocytes, monocytes and neutrophils (Extended Data Fig. 2a–c), with increased circulating activated CD8$^+$ T cells (Extended Data Fig. 2b, c), and augmented immunohistological staining of CD3 and Ki67 in the spleen (Extended Data Fig. 2d). Notably, LAP-deficient animals also contained increased serum levels of anti-double-stranded DNA (dsDNA) antibodies and anti-nuclear antibodies (Fig. 1b, c), as well as a broad array of antibodies against autoantigens commonly associated with SLE (Fig. 1d and Extended Data Fig. 3). LAP-deficient animals also presented with IgG and complement C1q deposition in the glomeruli of kidneys (Fig. 2a–d, Extended Data Fig. 4a, b). In addition, LAP-deficient animals displayed indications of kidney damage[15], and exhibited increased functional markers of kidney injury, such as increased serum creatinine (Fig. 2e), blood urea nitrogen, and proteinuria (Extended Data Fig. 4c, d). Histologically, kidneys from aged LAP-deficient animals displayed endocapillary proliferative glomerulonephritis (Extended Data Fig. 4e). Increased expression of type I interferon (IFN)-regulated genes, termed the IFN signature, has been reported in SLE patients[16]. Analysis revealed increased expression of IFN signature genes, such as *Ddx58* (which encodes RIG-I) and *Isg95* (also known as *Cmtr1*), in the spleens of aged LAP-deficient animals (Extended Data Fig. 5a). By contrast, none of these pathologies was observed in animals lacking autophagy components dispensable for LAP (Fig. 2a–e and Extended Data Figs 4a–e, 5a). Collectively, these observations suggest that LAP deficiency, but not autophagy deficiency, causes an autoinflammatory, lupus-like syndrome in mice.

The kinetics of disease we observed in all LAP-deficient animals was markedly similar to that of animals lacking T-cell immunoglobulin mucin protein 4 (TIM4) (Figs 1a, b and 2a, e). TIM4 is required for engulfment of dying cells in several macrophage populations, and animals lacking TIM4 display lupus-like disease[2], as do animals defective for other proteins involved in the clearance of dying cells, including MERTK, MFG-E8, and C1q (ref. 1). However, we found that neither bone-marrow-derived macrophages (Extended Data Fig. 5b) nor peritoneal exudate macrophages from 52-week-old mice of any genotype (Extended Data Fig. 5c) showed any defects in the engulfment of dying cells *in vitro*. We therefore examined the role of LAP in the response

[1]Department of Immunology, St Jude Children's Research Hospital, Memphis, Tennessee 38105, USA. [2]Immunity, Inflammation, and Disease Laboratory, National Institute of Environmental Health Sciences, 111 T.W. Alexander Drive, Research Triangle Park, North Carolina 27709, USA. [3]Department of Pathology and Immunology, Washington University School of Medicine, St Louis, Missouri 63110, USA. [4]University of Texas Southwestern Medical Center, Dallas, Texas 75390, USA. [5]Division of Nephrology and Hypertension, Christian-Albrechts-University, Kiel 24105, Germany.

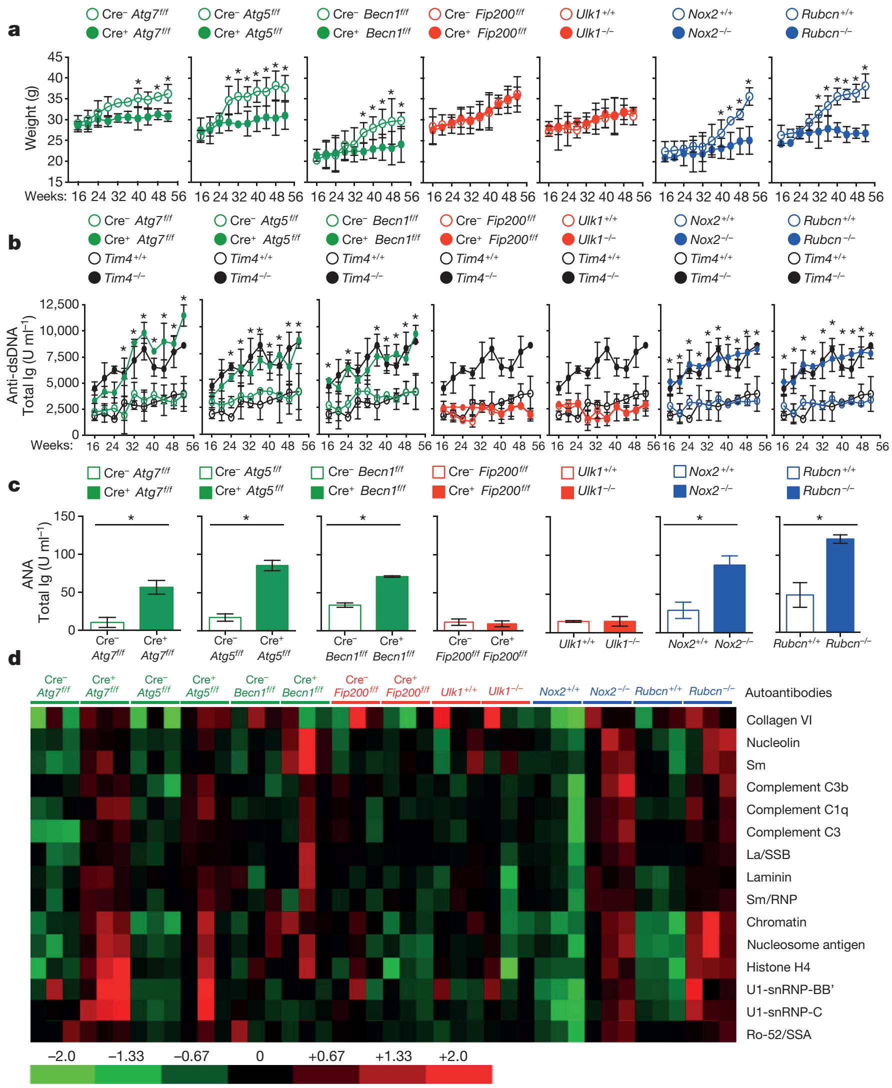



**Figure 1 | Mice with LAP deficiencies display symptoms of autoinflammatory disorder.** Wild-type and deficient littermates were co-housed and aged for 52 weeks at St Jude Children's Research Hospital (SJCRH). **a**, Weights. **b**, Anti-dsDNA antibodies (total Ig). **c**, **d**, Anti-nuclear antigens (ANA; total Ig) in animals aged 52 weeks (**c**). Antibodies to autoantigens commonly associated with autoimmune and autoinflammatory disorders. Three mice per genotype, normalized background signals (**d**). In all cases, Cre indicates LysM-Cre. Error bars represent s.d. $*P < 0.001$ (Student's *t*-test). Animal numbers are provided in the Methods. Colour scheme represents LAP-deficient, autophagy-deficient genotypes (green), autophagy-deficient, LAP-sufficient (red), and autophagy-sufficient, LAP-deficient (blue). Values for one cohort of $Tim4^{+/+}$ and $Tim4^{-/-}$ animals are shown for comparison in all cases (black) in **a** and **b**.

to dying cells *in vivo*. PKH26-labelled wild-type C57Bl/6 thymocytes were ultraviolet (UV)-irradiated to trigger apoptosis and immediately injected into wild-type animals, or animals with LysM-Cre-mediated deficiency of ATG7 (LAP-deficient, autophagy-deficient), LysM-Cre-mediated deficiency of FIP200 (LAP-sufficient, autophagy-deficient), or ubiquitous deletion of RUBCN (LAP-deficient, autophagy-sufficient), all of which also expressed transgenic green fluorescent protein (GFP)-tagged LC3 (ref. 5). Clearance of dying thymocytes and induction of LC3-II (a measure of LC3 conversion[5]) were monitored in spleen, liver and kidney. While both wild-type and animals with FIP200-deficiency effectively cleared dying cells (Fig. 3a, b and Extended Data Fig. 6a) and converted GFP–LC3 (Extended Data Fig. 6b–d), animals with ATG7- or RUBCN-deficiency did not, despite engulfment (Fig. 3a, b and Extended Data Fig. 6a, b, d, e). These data are consistent with our observations *in vitro*[4], and support the conclusion that LAP is required for effective degradation of engulfed, dying cells *in vivo*. Dying cells were engulfed by CD11b$^{+}$ F4/80$^{+}$ macrophages, CD11b$^{+}$ Gr1$^{+}$ granulocytes, CD11b$^{+}$ CD115$^{+}$ monocytes, and CD11b$^{+}$ CD11c$^{+}$ dendritic cells, equivalently in wild-type and $Rubcn^{-/-}$ mice, but not in $Tim4^{-/-}$ (also known as $Timd4^{-/-}$) mice (Extended Data Fig. 6e). However, while the frequency of engulfment declined by 48 h in all cellular subsets in wild-type mice, they remained elevated in $Rubcn^{-/-}$ mice, consistent with a failure of a LAP-dependent mechanism to degrade engulfed corpses (Extended Data Fig. 6e).

Previously, we had found that in contrast to wild-type or $Ulk1^{-/-}$ macrophages, Cre$^{+}$ $Atg7^{f/f}$ macrophages produce increased levels of inflammatory cytokines, such as IL-1β and IL-6 *in vitro*[4]. We therefore examined cytokine production after ingestion of dying cells in macrophages lacking different components of the LAP or autophagy pathways (Extended Data Fig. 7a–d). LAP-deficient (Cre$^{+}$ $Atg7^{f/f}$, Cre$^{+}$ $Becn1^{f/f}$, Cre$^{+}$ $Atg3^{f/f}$, $Nox2^{-/-}$ (also known as $Cybb^{-/-}$) and $Rubcn^{-/-}$; Cre indicates LysM-Cre throughout) but not LAP-sufficient (Cre$^{+}$ $Fip200^{f/f}$, Cre$^{+}$ $Atg14^{f/f}$) macrophages produced IL-1β, IL-6 and IP-10 (also known as CXCL10), upon engulfment of dying cells (Extended Data Fig. 7a–c). Conversely, LAP-sufficient, but not LAP-deficient macrophages produced IL-10 upon engulfment (Extended Data Fig. 7d). We then examined the effects of dying cells on serum cytokine production *in vivo*, after injection of UV-irradiated thymocytes (Fig. 3c, d). Notably, serum IL-1β, IL-6 and MIP-1β (also known as CCL4) were acutely increased in LAP-deficient animals (ATG7 or RUBCN), but not in LAP-sufficient animals (wild-type or FIP200) (Fig. 3c, d). As observed *in vitro*, LAP-sufficient animals produced increased serum IL-10 in response to dying cells, whereas LAP-deficient animals did not (Fig. 3c, d). Therefore, LAP, but not canonical autophagy, is required for the production of IL-10 in response to apoptotic cell engulfment, and LAP suppresses the production of inflammatory cytokines under these conditions.

We next asked whether repeated injection of apoptotic thymocytes into LAP-deficient animals could exacerbate the SLE-like phenotype observed in aged LAP-deficient animals. Beginning at 6 weeks of age, $Rubcn^{+/+}$ and $Rubcn^{-/-}$ animals were injected with UV-irradiated thymocytes over an 8-week period. Uninjected $Rubcn^{+/+}$ animals showed a minimal

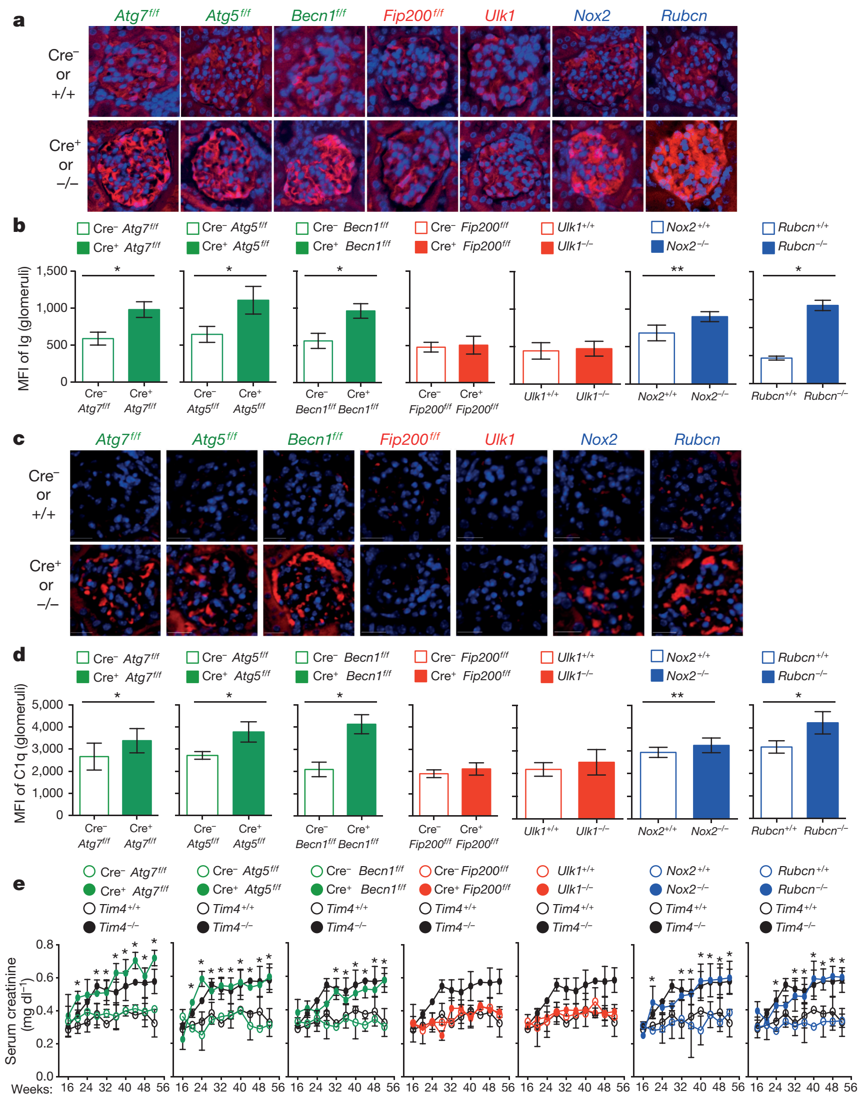



**Figure 2 | Mice with LAP deficiencies display kidney pathology. a–d**, Appearance of kidneys of co-housed, 52-week-old animals. DAPI (blue), anti-IgG (red, **a**), anti-C1q (red, **c**). Original magnifications, ×100. Mean fluorescent intensity (MFI) of anti-IgG (**b**) and anti-C1q (**d**) in glomeruli. **e**, Serum creatinine. Animal numbers are provided in Methods. Error bars represent s.d. (**P* < 0.001, ***P* < 0.05, Student's *t* test). For histological assessment, at least 15 glomeruli were evaluated for each genotype. Colour scheme represents LAP-deficient, autophagy-deficient genotypes (green), autophagy-deficient, LAP-sufficient (red), and autophagy-sufficient, LAP-deficient (blue). Values for one cohort of $Tim4^{+/+}$ and $Tim4^{-/-}$ animals are shown for comparison in all cases (black) in **e**.

increase in anti-nuclear antigens (ANA) and anti-dsDNA autoantibodies after 8 weeks, and no increase attributable to injection of dying cells. $Rubcn^{-/-}$ animals, however, displayed a significant increase in serum levels of ANA and anti-dsDNA autoantibodies after 8 weeks of dying cell injections, above pre-injection and age-matched, uninjected controls (Fig. 3e). Furthermore, these animals displayed IgG and C1q deposition in the glomeruli of kidneys (Extended Data Fig. 8a, b), and injected $Rubcn^{-/-}$ animals displayed increased levels of alanine aminotransferase, indicative of tissue damage (Extended Data Fig. 8c). Collectively, these data demonstrate that defective dead cell clearance associated with LAP deficiency can result in development of SLE-like disease.

We next examined spontaneous levels of serum cytokines with age in animals with or without LAP. All genotypes lacking LAP (Cre$^{+}$ *Atg7*$^{f/f}$, Cre$^{+}$ *Atg5*$^{f/f}$, Cre$^{+}$ *Becn1*$^{f/f}$, $Nox2^{-/-}$ and $Rubcn^{-/-}$) displayed increased levels of IL-1β, IL-6, IL-12p40 and IP-10 (Fig. 4a–d), as well as KC (also known as CXCL1), MIP-1β and MCP-1 (also known as CCL2) (Extended Data Fig. 8d–f). Wild-type animals and animals lacking canonical autophagy, but not LAP (in monocytes or systemically), did not display increased inflammatory cytokines at any time point (Fig. 4a–d and Extended Data Fig. 8d–f). By contrast, serum IL-10 levels, which increased with age in LAP-sufficient strains, were undetectable in animals lacking LAP (Fig. 4e). The patterns and kinetics of cytokine levels were similar to that observed in $Tim4^{-/-}$ animals (Fig. 4a–e and Extended Data Fig. 8d–f).

Our observations indicated that defects in LAP, but not canonical autophagy, cause an autoinflammatory, lupus-like syndrome in mice. To test this idea further, we examined both LAP-sufficient and LAP-deficient mice bred in an independent facility. Mice with ATG5- or ATG3-deficient myeloid cells (defective in LAP and autophagy) displayed increased levels of IL-1β, IL-6, IL-12p40, IP-10, KC, MIP-1β and MCP-1 at 52 weeks of age (Extended Data Fig. 9a–g). These LAP-deficient animals also displayed significantly lower levels of IL-10 than controls (Extended Data Fig. 9h). Furthermore, LAP-deficient animals displayed elevated anti-dsDNA antibodies (Extended Data Fig. 9i) and serum creatinine (Extended Data Fig. 9J). LAP-deficient animals also contained a broad array of antibodies against auto-antigens commonly associated with SLE (Extended Data Fig. 10). Of note, none of these effects was observed in animals with ATG14- or FIP200-deficiency (defective autophagy but normal LAP[3,6,7,11,13]) (Extended Data Figs 9a–j, 10). It is noteworthy that these effects in two different facilities were observed in C57Bl/6 background animals, which are generally resistant to lupus-like disease[17]. Previous studies have shown that one of the LAP-deficient genotypes, $Nox2^{-/-}$, causes accelerated, severe lupus-like disease when bred on the lupus-prone MRL.lpr background[18].

Altogether, these data suggest that defective LAP results in a failure to digest engulfed dying cells, leading to increased inflammatory cytokine production and a lupus-like syndrome. In another study, animals in which lung macrophages were incapable of engulfment owing to deletion of RAC1 were sensitive to inflammatory cytokine production and inflammatory disease after introduction of dying cells into the lung[19]. Similarly, TIM4-deficient mice, which exhibit defective dead cell

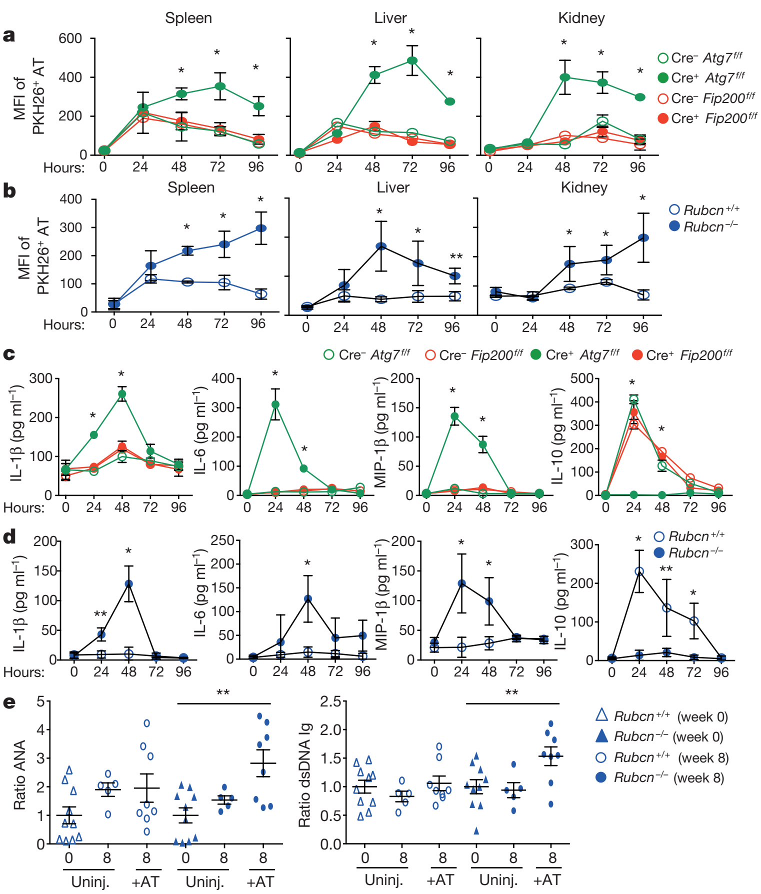



**Figure 3 | Mice with LAP deficiencies display defective clearance of engulfed, dying cells, resulting in increased production of pro-inflammatory cytokines. a–d,** $1 \times 10^7$ PKH26-labelled UV-irradiated wild-type thymocytes were injected intravenously into indicated animals expressing GFP–LC3. Apoptotic thymocytes (AT) in spleen, liver and kidney of indicated animals measured by flow cytometry (**a**, **b**). Indicated serum cytokines (**c**, **d**). Error bars represent s.d. ($n = 4$, $*P < 0.001$, $**P < 0.05$, Student's $t$-test). **e**, $2 \times 10^7$ UV-irradiated wild-type thymocytes were injected intravenously six times over 8 weeks into indicated animals (aged 6 weeks). Serum anti-nuclear antibodies (total Ig) and anti-dsDNA antibodies (total Ig) are shown at 16 weeks from uninjected (uninj.) and injected (+AT) animals. Results are presented as ratio to average value before injection for each individual animal. Error bars represent s.e.m. ($n = 4$, $**P < 0.05$, Student's $t$-test). The colour scheme represents LAP-deficient, autophagy-deficient genotypes (green), autophagy-deficient, LAP-sufficient (red), and autophagy-sufficient, LAP-deficient (blue).

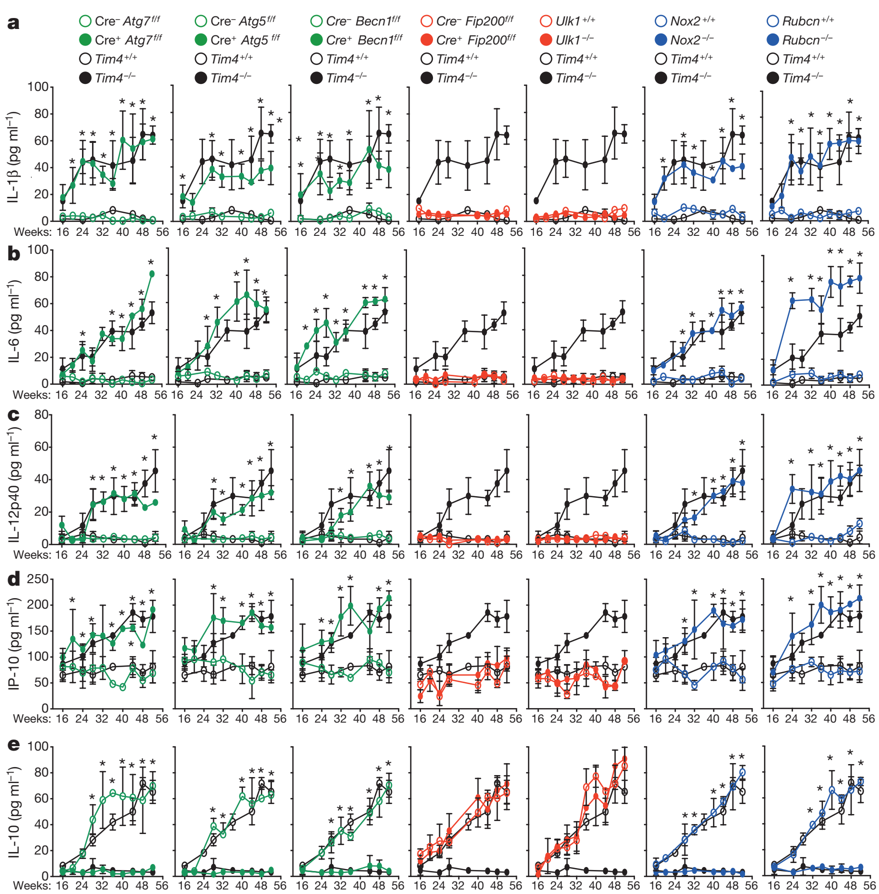



**Figure 4 | Mice with LAP deficiencies display symptoms of an autoinflammatory disorder. a–e**, Indicated serum cytokines in co-housed 52-week-old animals. Error bars represent s.d. ($*P < 0.001$, Student's $t$-test). Numbers of animals are provided in Methods. Colour scheme represents LAP-deficient, autophagy-deficient genotypes (green), autophagy-deficient, LAP-sufficient (red), and autophagy-sufficient, LAP-deficient (blue). Values for one cohort of $Tim4^{+/+}$ and $Tim4^{-/-}$ animals are shown for comparison in all cases (black) in **a–e**.

engulfment[2], showed spontaneous increases in serum inflammatory cytokines with age (Fig. 4 and Extended Data Fig. 8) as well as lupus-like disease (Figs 1 and 2). By contrast, macrophages defective for LAP engulf dying cells, but fail to digest them efficiently[4,7]. This suggests that LAP-dependent digestion of dying cells, rather than engulfment alone, suppresses an inflammatory response by macrophages. In the absence of LAP (lack of BECN1, ATG7, ATG5, NOX2, RUBCN), macrophages engulf dying cells and produce inflammatory cytokines, and animals manifest lupus-like disease. However, when canonical autophagy, but not LAP, is defective (lack of FIP200, ULK1), dying cells are engulfed, macrophages produce IL-10 but not inflammatory cytokines, and no lupus-like disease is observed.

MRL.lpr mice lacking IL-10 display markedly accelerated lupus-like disease[20]. While macrophages, monocytes and B cells are the major source of IL-10, specific deletion of IL-10 in B cells had no effect on pathogenesis in MRL.lpr mice[21]. Notably, one study found that injection of dendritic cells that had engulfed necrotic cells into IL-10-deficient, but not wild-type mice induced a pronounced lupus-like disease[22]. Thus, the role of LAP in the production of IL-10 may contribute to the disease effects we observed. However, most studies have implicated increased IL-10 levels in mouse and human SLE[20,23,24], perhaps involved with the activation of B lymphocytes[25]. While IL-10 production in response to dying cells was compromised in LAP-deficient macrophages, the production of IL-10 in response to other stimuli may remain intact, and thus increased IL-10 in SLE may be due to other events in the pathogenesis of SLE.

Genome-wide association studies have implicated autophagy in SLE (*Atg5* (refs 6, 8) and possibly *Atg7* (ref. 9)) and in Crohn's disease (*Atg16l*; ref. 26). It is notable in this context that the ATG5 association with SLE may depend on polymorphisms in IL-10 (refs 8, 27). Other studies have suggested that autophagy suppresses the inflammasome[28], providing a possible link between autophagy and inflammatory disease. However, the autophagic components identified in these studies are also required for LAP. Furthermore, mice[18] and humans[29] lacking NOX2 develop SLE, and our studies suggest that defective LAP in this context may contribute to this effect. Our findings implicate a non-canonical autophagic process, LAP, in the control of inflammatory disease and suggest a link between the clearance of dying cells, autophagic processes, and inflammation in the control of SLE.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 6 March 2015; accepted 1 April 2016.**
**Published online 20 April; corrected online 4 May 2016**
**(see full-text HTML version for details).**

1. Ravichandran, K. S. Find-me and eat-me signals in apoptotic cell clearance: progress and conundrums. *J. Exp. Med.* **207,** 1807–1817 (2010).
2. Rodriguez-Manzanet, R. *et al*. T and B cell hyperactivity and autoimmunity associated with niche-specific defects in apoptotic body clearance in TIM-4-deficient mice. *Proc. Natl Acad. Sci. USA* **107,** 8706–8711 (2010).
3. Kim, J. Y. *et al.* Noncanonical autophagy promotes the visual cycle. *Cell* **154,** 365–376 (2013).
4. Martinez, J. *et al.* Microtubule-associated protein 1 light chain 3 $\alpha$ (LC3)-associated phagocytosis is required for the efficient clearance of dead cells. *Proc. Natl Acad. Sci. USA* **108,** 17396–17401 (2011).
5. Martinez, J. *et al.* Molecular characterization of LC3-associated phagocytosis (LAP) reveals distinct roles for Rubicon, NOX2, and autophagy proteins. *Nature Cell Biol.* **17,** 893–906 (2015).
6. Henault, J. *et al.* Noncanonical autophagy is required for type I interferon secretion in response to DNA-immune complexes. *Immunity* **37,** 986–997 (2012).
7. Florey, O., Kim, S. E., Sandoval, C. P., Haynes, C. M. & Overholtzer, M. Autophagy machinery mediates macroendocytic processing and entotic cell death by targeting single membranes. *Nature Cell Biol.* **13,** 1335–1343 (2011).
8. Zhou, X. J. *et al.* Genetic association of PRDM1–ATG5 intergenic region and autophagy with systemic lupus erythematosus in a Chinese population. *Ann. Rheum. Dis.* **70,** 1330–1337 (2011).
9. Clarke, A. J. *et al.* Autophagy is activated in systemic lupus erythematosus and required for plasmablast development. *Ann. Rheum. Dis.* **74,** 912–920 (2015).
10. Sanjuan, M. A. *et al.* Toll-like receptor signalling in macrophages links the autophagy pathway to phagocytosis. *Nature* **450,** 1253–1257 (2007).
11. Gan, B. & Guan, J. L. FIP200, a key signaling node to coordinately regulate various cellular processes. *Cell. Signal.* **20,** 787–794 (2008).
12. Mizushima, N. & Levine, B. Autophagy in mammalian development and differentiation. *Nature Cell Biol.* **12,** 823–830 (2010).
13. Matsunaga, K. *et al.* Two Beclin 1-binding proteins, Atg14L and Rubicon, reciprocally regulate autophagy at different stages. *Nature Cell Biol.* **11,** 385–396 (2009).
14. Clausen, B. E., Burkhardt, C., Reith, W., Renkawitz, R. & Forster, I. Conditional gene targeting in macrophages and granulocytes using LysMcre mice. *Transgenic Res.* **8,** 265–277 (1999).
15. Theofilopoulos, A. N. & Dixon, F. J. Murine models of systemic lupus erythematosus. *Adv. Immunol.* **37,** 269–390 (1985).
16. Rönnblom, L. & Eloranta, M. L. The interferon signature in autoimmune diseases. *Curr. Opin. Rheumatol.* **25,** 248–253 (2013).
17. Morel, L., Rudofsky, U. H., Longmate, J. A., Schiffenbauer, J. & Wakeland, E. K. Polygenic control of susceptibility to murine systemic lupus erythematosus. *Immunity* **1,** 219–229 (1994).
18. Campbell, A. M., Kashgarian, M. & Shlomchik, M. J. NADPH oxidase inhibits the pathogenesis of systemic lupus erythematosus. *Sci. Transl. Med.* **4,** 157ra141 (2012).
19. Juncadella, I. J. *et al.* Apoptotic cell clearance by bronchial epithelial cells critically influences airway inflammation. *Nature* **493,** 547–551 (2013).
20. Yin, Z. *et al.* IL-10 regulates murine lupus. *J. Immunol.* **169,** 2148–2155 (2002).
21. Teichmann, L. L. *et al.* B cell-derived IL-10 does not regulate spontaneous systemic autoimmunity in MRL.*Fas$^{lpr}$* mice. *J. Immunol.* **188,** 678–685 (2012).
22. Ma, L. *et al.* Systemic autoimmune disease induced by dendritic cells that have captured necrotic but not apoptotic cells in susceptible mouse strains. *Eur. J. Immunol.* **35,** 3364–3375 (2005).
23. Perry, D., Sang, A., Yin, Y., Zheng, Y. Y. & Morel, L. Murine models of systemic lupus erythematosus. *J. Biomed. Biotechnol.* **2011,** 271694 (2011).
24. Houssiau, F. A. *et al.* Serum interleukin 10 titers in systemic lupus erythematosus reflect disease activity. *Lupus* **4,** 393–395 (1995).
25. Blenman, K. R. *et al.* IL-10 regulation of lupus in the NZM2410 murine model. *Lab. Invest.* **86,** 1136–1148 (2006).
26. Hampe, J. *et al.* A genome-wide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crohn disease in *ATG16L1*. *Nature Genet.* **39,** 207–211 (2007).
27. López, P., Alonso-Pérez, E., Rodriguez-Carrio, J. & Suárez, A. Influence of Atg5 mutation in SLE depends on functional IL-10 genotype. *PLoS ONE* **8,** e78756 (2013).
28. Shi, C. S. *et al.* Activation of autophagy by inflammatory signals limits IL-1beta production by targeting ubiquitinated inflammasomes for destruction. *Nature Immunol.* **13,** 255–263 (2012).
29. De Ravin, S. S. *et al.* Chronic granulomatous disease as a risk factor for autoimmune disease. *J. Allergy Clin. Immunol.* **122,** 1097–1103 (2008).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** The authors thank T. Brewer, P. Fitzgerald, M. Henderson, J. Kolb and T. Oguin for technical assistance. We also thank K. Gerrish and R. Fannin for their work and analysis of the Nanostring data. This work was supported by the Intramural Research Program of the National Institutes of Health, NIEHS (1ZIAES10328601), as well as grants from the US National Institutes of Health (RO1 AI40646, U19 AI109725), the Lupus Research Institute, the German Research Foundation (EXC306), and ALSAC.

**Author Contributions** J.M. and D.R.G. designed the experiments; J.M. performed and analysed the experiments; L.D.C., S.P., M.Y., Q.L., Q.-Z.L., M.Y., L.J., C.G., A.L. and R.O. performed and analysed specific experiments; J.M., H.W.V. and D.R.G. wrote the manuscript.

**Author Information** Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to D.R.G. (douglas.green@stjude.org) or J.M. (jennifer.martinez3@nih.gov).

# METHODS

**Mice and primary cells.** All mice were housed specific pathogen-free.

*Ulk1*$^{-/-}$ mice were provided by M. Kundu. *Atg7*$^{f/f}$ mice (provided by M. Komatsu) were bred to LysM-Cre$^{+}$ mice (provided by P. Murray) and GFP-LC3$^{+}$ mice to generate LysM-Cre$^{+}$ *Atg7*$^{f/f}$ GFP–LC3$^{+}$ versions of these strains. *Nox2*$^{-/-}$ mice were purchased from Jackson Laboratories. LysM-Cre$^{+}$ *Becn1*$^{f/f}$ (E. Rucker), LysM-Cre$^{+}$ *Atg5*$^{f/f}$ (T. A. Ferguson), and LysM-Cre$^{+}$ *Fip200*$^{f/f}$ (J.-L. Guan) were bred to GFP-LC3$^{+}$ mice to generate GFP–LC3$^{+}$ versions of these strains. *Tim4*$^{-/-}$ mice were provided by V. Kuchroo. *Rubcn*$^{+/+}$ and *Rubcn*$^{-/-}$ mice were generated using CRISPR/Cas9 gene editing technology[5]. LysM-Cre$^{+}$ *Atg14*$^{f/f}$, LysM-Cre$^{+}$ *Atg3*$^{f/f}$, LysM-Cre$^{+}$ *Atg5*$^{f/f}$, LysM-Cre$^{+}$ *Fip200*$^{f/f}$ mice (and control littermates) were bred and maintained in the Washington University facility. The St Jude Institutional Animal Care and Use Committee approved all procedures in accordance with the Guide for the Care and Use of Animals.

Bone-marrow-derived macrophages were generated from bone marrow progenitors obtained from littermates. Freshly prepared bone marrow cells were cultured in DMEM medium supplemented with 10% heat-inactivated FCS, 2 mM L-glutamine, 10 mM HEPES buffer, 50 μg ml$^{-1}$ penicillin, and non-essential amino acids in the presence of 20 ng ml$^{-1}$ rmM-CSF (Peprotech) for 6 days. Nonadherent cells were removed on day 6, and adherent macrophages were detached from plates and re-plated for experimental use.

**Ageing studies.** Male wild-type and knockout littermates were co-housed and allowed to age for 52 weeks. Animals were weighed and bled retro-orbitally monthly, and serum was collected for use in assays (below). Numbers of animals were as follows (in all cases, Cre indicates LysM-Cre). Studies conducted at St Jude Children's Research Hospital and reported in Figs 1, 2, 4 and Extended Data Figs 2–5 and 8: Cre$^{-}$ and Cre$^{+}$ *Atg7*$^{f/f}$, $n=24$ per genotype; Cre$^{-}$ and Cre$^{+}$ *Atg5*$^{f/f}$, $n=14$ per genotype; Cre$^{-}$ and Cre$^{+}$ *Becn1*$^{f/f}$, $n=20$ per genotype; Cre$^{-}$ and Cre$^{+}$ *Fip200*$^{f/f}$, $n=16$ per genotype; *Ulk1*$^{+/+}$ and *Ulk1*$^{-/-}$, $n=14$ per genotype; *Nox2*$^{+/+}$ and *Nox2*$^{-/-}$, $n=10$ per genotype; *Rubcn*$^{+/+}$ and *Rubcn*$^{-/-}$, $n=14$ per genotype. Studies conducted at Washington University and reported in Extended Data Figs 9 and 10: Cre$^{-}$ and Cre$^{+}$ *Atg5*$^{f/f}$, $n=5$ per genotype; Cre$^{-}$ and Cre$^{+}$ *Atg3*$^{f/f}$, $n=4$ per genotype; Cre$^{-}$ and Cre$^{+}$ *Fip200*$^{f/f}$, $n=4$ per genotype; Cre$^{-}$ and Cre$^{+}$ *Atg14*$^{f/f}$, $n=4$ per genotype.

**Induction of apoptosis in thymocytes.** Apoptosis was induced in wild-type C57Bl/6 thymocytes by UV irradiation (20 J m$^{-2}$). Thymocytes were washed twice with PBS before experimental use.

**Staining of apoptotic thymocytes and *in vivo* adoptive transfer of labelled apoptotic thymocytes.** UV-treated thymocytes were stained with 20 M PKH26 Red (Sigma), per manufacturer's instructions. $1\times10^{7}$ PKH26-labelled, apoptotic thymocytes were injected intravenously into GFP–LC3$^{+}$ animals, and serum, kidney, liver and spleen was collected at 0, 24, 48, 72 and 96 h after injection. Kidney sections were analysed for persistence of PKH26-labelled apoptotic cells using the Nikon800 microscope. Kidney, liver and spleen samples were analysed for PKH26-labelled apoptotic cells using flow cytometry. Additionally, samples were washed once with FACS buffer and permeabilized with digitonin (Sigma, 200 μg ml$^{-1}$) for 15 min on ice. Cells were then washed three times with FACS buffer and analysed by flow cytometry for membrane-bound GFP–LC3-II associated with engulfed PKH26-labelled thymocytes. For quantification of phagocytosis, spleens were harvested and stained for fluorescently conjugated surface markers for macrophages (CD11b$^{+}$ F4/80$^{+}$), neutrophils (CD11b$^{+}$ Gr-1$^{+}$), monocytes (CD11b$^{+}$ CD115$^{+}$), and dendritic cells (CD11b$^{+}$ CD11c$^{+}$). Phagocytic efficiency of each cell type (singlets/cell surface markers$^{+}$/PKH26$^{+}$) was quantified by flow cytometry (percentage PKH26).

**Repeated injection of apoptotic thymocytes.** Six-week-old *Rubcn*$^{+/+}$ and *Rubcn*$^{-/-}$ littermates were used. Serum was collected from all animals before injection (week 0). $2.0\times10^{7}$ UV-irradiated thymocytes (20 J m$^{-2}$) suspended in sterile phosphate buffer were injected intravenously (i.v.) into anaesthesized mice, once a week for four consecutive weeks (from weeks 1 to 4). After a resting period of 15 days, the injections were resumed and carried out for other 2 weeks (weeks 6 and 7). Serum was collected 1 week after the last injection (week 8) and assessed for levels of anti-dsDNA autoantibodies (total Ig), ANA (total Ig), and alanine aminotransferase. At week 8, mice were euthanized, the kidneys were collected and stained for immunofluorescence (below).

**Collection and co-culture of peritoneal exudate cells.** For collection of peritoneal exudate cells, mice were injected intraperitoneally (i.p.) with 2 ml of 3% Brewer's thioglycollate and euthanized 96 h later. The peritoneum was washed with 10 ml ice-cold PBS three times. Cells were centrifuged (225*g*, 6 min, 4 °C) and washed twice with sterile PBS. Peritoneal exudate cells were resuspended in DMEM plus 10% FBS, counted and plated at $5\times10^{5}$ cells per well in a 12-well plate. Cells were allowed to settle for 2 h (37 °C, 5% $CO_2$) before co-culture with UV-irradiated wild-type thymocytes.

**Effects of dying cells on macrophages *in vitro*.** Apoptotic thymocytes were added to BMDM cultures at a ratio of 10:1 (dead cell:macrophage). Supernatant was collected after 24 h of culture and analysed for cytokines (see below).

**Flow cytometry analysis.** Spleens, livers and kidneys were collected from animals at the indicated time points, and single-cell suspensions were generated. Cells were washed once with FACS buffer, and permeabilized with digitonin (Sigma, 200 μg ml$^{-1}$) for 15 min on ice. Cells were then washed three times with FACS buffer and analysed by flow cytometry for membrane-bound GFP–LC3-II. This assay removes the soluble, cytosolic form of GFP–LC3 (GFP–LC3-I), while the lipidated, membrane-bound GFP–LC3-II is retained, allowing total GFP fluorescence to be used as a measure of LC3-II generation, indicative of LAP. Permeabilized samples were first gated on singlets/PKH26$^{+}$, so as to determine the MFI of GFP–LC3-II associated with cells that had engulfed a PKH26$^{+}$ apoptotic thymocyte. For surface staining, blood, bone marrow or splenocytes were washed once with FACS buffer, incubated with Fc Block and stained with the indicated fluorescent antibodies (Biolegend) on ice for 20 min. Cells were then washed twice with FACS buffer and analysed by flow cytometry. Data were acquired using an LSRII cytometer (BD).

**Quantification of phagocytosis.** Phagocytosis was quantified using flow cytometry analysis (described above). Apoptotic thymocytes were stained with CellTrace Violet (Molecular Probes) or PKH26 (Sigma-Aldrich) per manufacturer's protocol. Percentage phagocytosis equals the percentage of cells that have engulfed CellTrace Violet$^{+}$ or PKH26$^{+}$ apoptotic thymocytes.

**Immunofluorescent staining and analysis of IgG and C1q deposition in kidney sections.** Kidneys were collected from animals at 32, 52 or 8 weeks after chronic apoptotic thymocyte injection (above). Organs were sectioned and mounted on slides. Slides were fixed with 4% formaldehyde for 20 min at 4 °C. Following fixation, slides were blocked and permeabilized in block buffer (1% BSA, 0.1% Triton in PBS) for 1 h at room temperature. Slides were washed extensively in TBS containing 0.05% Tween-20 (TBS-Tween), incubated with Alexa-Fluor 647-conjugated anti-IgG (Invitrogen) for 1 h at room temperature, and mounted with VectaShield with DAPI (Vector Labs). Alternatively, slides were washed extensively in TBS-Tween, incubated with anti-C1q (clone 4.8, Abcam) for 1 h at room temperature, washed again with TBS-Tween, incubated with Cy3-conjugated donkey anti-rabbit IgG (Jackson ImmunoResearch) and Alexa-Fluor 488-conjugated wheat germ agglutinin (Molecular Probes) for 1 h at room temperature, and mounted with VectaShield with DAPI (Vector Labs). Images were analysed using an Olympus BX51 FL Microscope and Slidebook software. Masks were drawn around glomeruli, and the MFI values of anti-IgG or anti-C1q were calculated.

**Cytokine detection.** Supernatants were collected from macrophages fed with apoptotic thymocytes for 24 h. Cytokines released into supernatant were analysed by Luminex technologies (Millipore). Serum was collected from animals was analysed by Luminex technologies (Millipore).

**Detection of serum creatinine.** The Veterinary Pathology Core at St Jude Children's Research Hospital measured serum creatinine.

**Blood and urine clinical chemistry.** The Veterinary Pathology Core at St Jude Children's Research Hospital assessed differential blood counts, alanine aminotransferase, and proteinuria (albumin to creatinine ratio). The Clinical Pathology Core at the National Institute of Environmental Health Sciences performed blood urea nitrogen analysis.

**Assessment of endocapillary proliferative glomerulonephritis.** Kidneys were collected from 52-week-old mice. Organs were sectioned, fixed in 10% formalin, and embedded in paraffin. Four to six micrometre serial sections were cut, deparaffinized, rehydrated and stained with haematoxylin and eosin. All slides were coded before evaluation, and only decoded upon collection of all data. Endocapillary proliferative glomerulonephritis, a glomerular disease pattern frequently associated with lupus nephritis, was assessed on a virtual scale ranging from 0 to 5, where '0' was considered 'indistinguishable compared to wild type control' and '5' was considered 'the maximal damage seen in all samples', based on the classification of glomerulonephritis in systemic lupus erythematosus. Features that influence this score are intraglomerular mesangial proliferation in relation to overall glomerular size, number of mesangial nuclei, intraluminal diameters of glomerular capillaries and the amount of mesangial matrix. Haematoxylin-and-eosin-stained sections were used to score at least 24 glomeruli in a maximum of 4 different specimens obtained from each group.

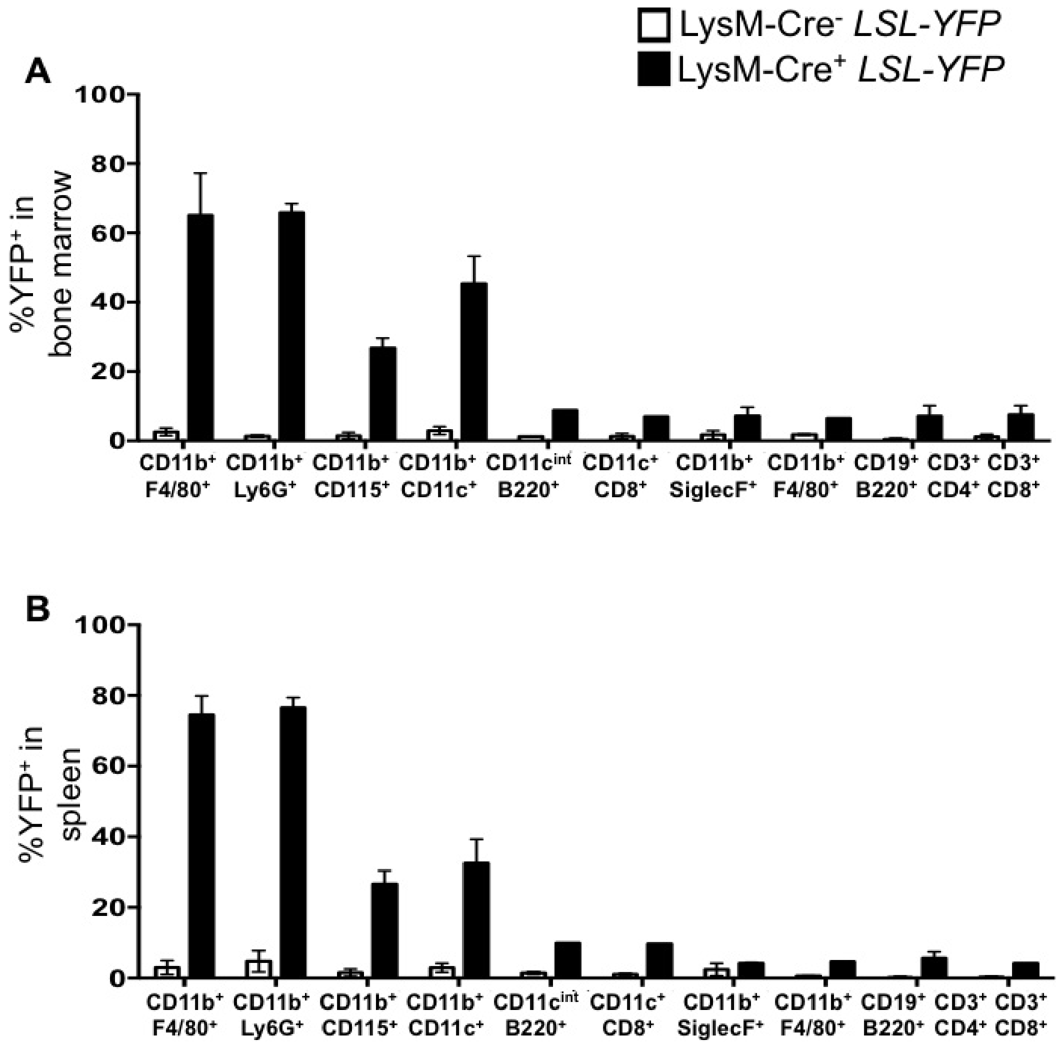
**Detection of anti-dsDNA antibodies and ANA.** The presence of anti-dsDNA antibodies in serum was tested using Mouse Anti-dsDNA Igs (Total A+G+M) ELISA Kit (Alpha Diagnostics International), per manufacturer's protocol. The presence of ANA in serum was tested using Mouse ANA/ENA Igs (Total A+G+M) ELISA Kit (Alpha Diagnostics International), per manufacturer's protocol.

**Detection of circulating autoantigen using autoantigen microarray.** Autoantibody reactivities against a penal of 124 autoantigens were measured

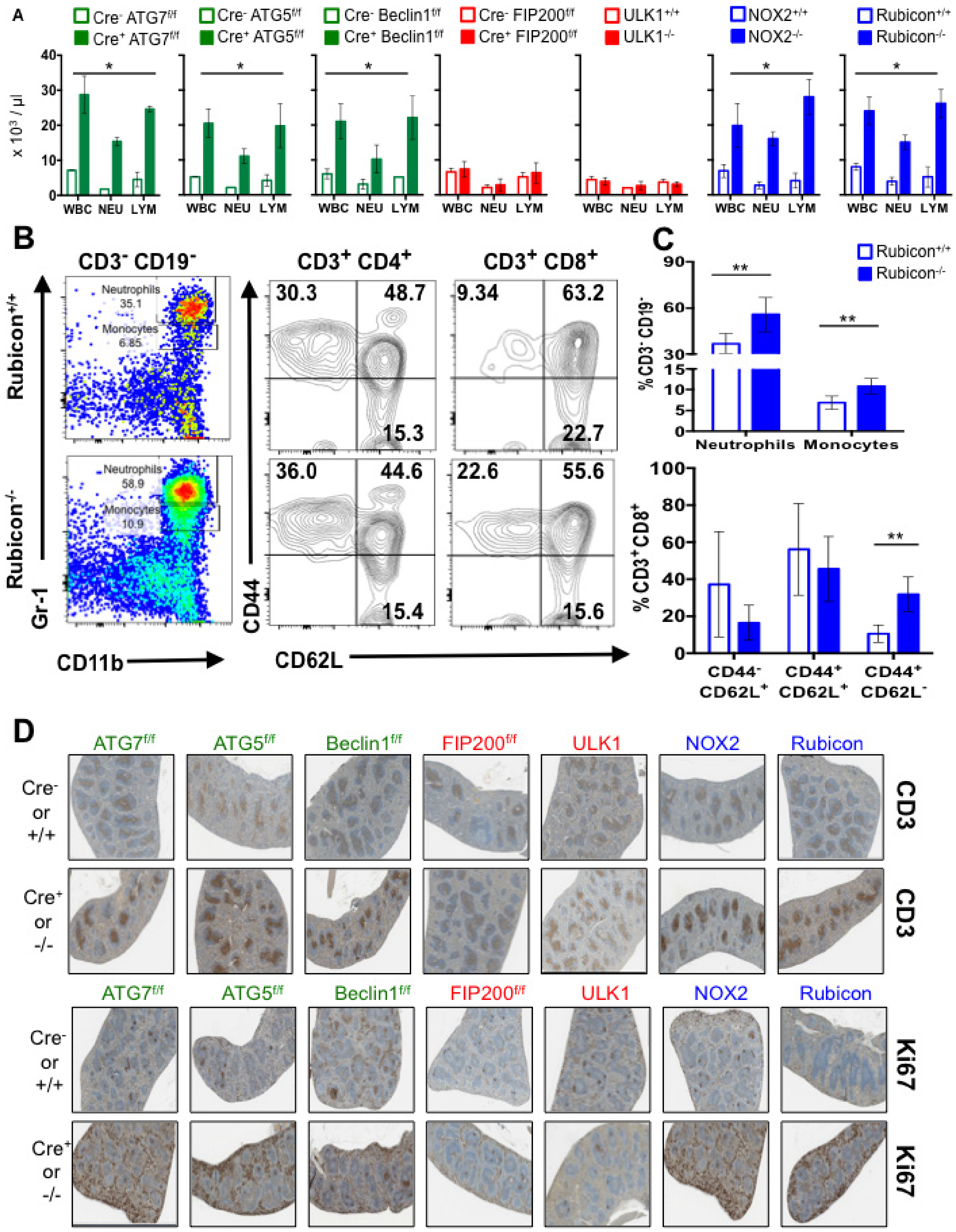
using an autoantigen microarray platform developed by University of Texas Southwestern Medical (https://microarray.swmed.edu/products/category/protein-array/). In brief, serum samples were pretreated with DNase-I and then diluted 1:50 in PBS plus 0.05% Tween-20 buffer for autoantibody profiling. The autoantigen array bearing 124 autoantigens and 4 control proteins were printed in duplicates onto Nitrocellulose film slides (Grace Bio-Labs). The diluted serum samples were incubated with the autoantigen arrays, and autoantibodies were detected with Cy3-labelled anti-mouse IgG and Cy5-labelled anti-mouse IgM using a Genepix 4200A scanner (Molecular Device) with laser wavelength of 532 nm and 635 nm. The resulting images were analysed using Genepix Pro 6.0 software (Molecular Devices). The median of the signal intensity for each spot was calculated and subtracted the local background around the spot, and data obtained from duplicate spots were averaged. The background subtracted signal intensity of each antigen was normalized to the average intensity of the total mouse IgG, which was included on the array as an internal control. Finally, the net fluorescence intensity for each antigen was calculated by subtracting a PBS control that was included for each experiment as negative control. The signal-to-noise ratio was used as a quantitative measurement of the true signal above background noise. Signal-to-noise ratio values equal to or greater than 3 were considered significantly higher than background, and therefore true signals. The net fluorescence intensity of each autoantibody was used to generate heatmaps using Cluster and Treeview software (http://bonsai.hgc.jp/~mdehoon/software/cluster/software.htm). Each row in the heatmap represents an autoantibody, and each column represents a sample. Red colour represents the signal intensity higher than the mean value of the raw, and green colour means signal intensity is lower than the mean value of the raw.

**RNA extraction and nanostring analysis.** Total RNA was isolated from the spleens from 52-week-old mice using NucleoSpin II kit (Macherey-Nagel) according to the manufacturer's instructions, and 50 ng was used to determine the absolute levels of gene expression. Hybridization and nCounter were performed according to the manufacturer's protocol (Nanostring Technologies). In brief, reactions were hybridized for 20 h at 65 °C, after which the products were used to run on the nCounter preparation station for removal of excess probes. Data were collected with the nCounter digital analyser by counting individual barcodes. Data generated from the nCounter digital analyser were examined with the nCounter digital analyser software system v2.1.1 (Nanostring Technologies). Data were normalized to the geometric means of spiked-in positive controls (controls for assay efficiency) and spiked-in negative controls (normalized for background). The data were further normalized to the housekeeping genes *Gapdh*, *Hprt* and *Tubb5* and are reported as normalized RNA counts (means ± s.e.m.). Nanostring RNA counts were analysed with the Partek Genomic Suite, to identify significantly regulated probe. Heatmaps of Nanostring data were generated with the Partek Genomic Suite.

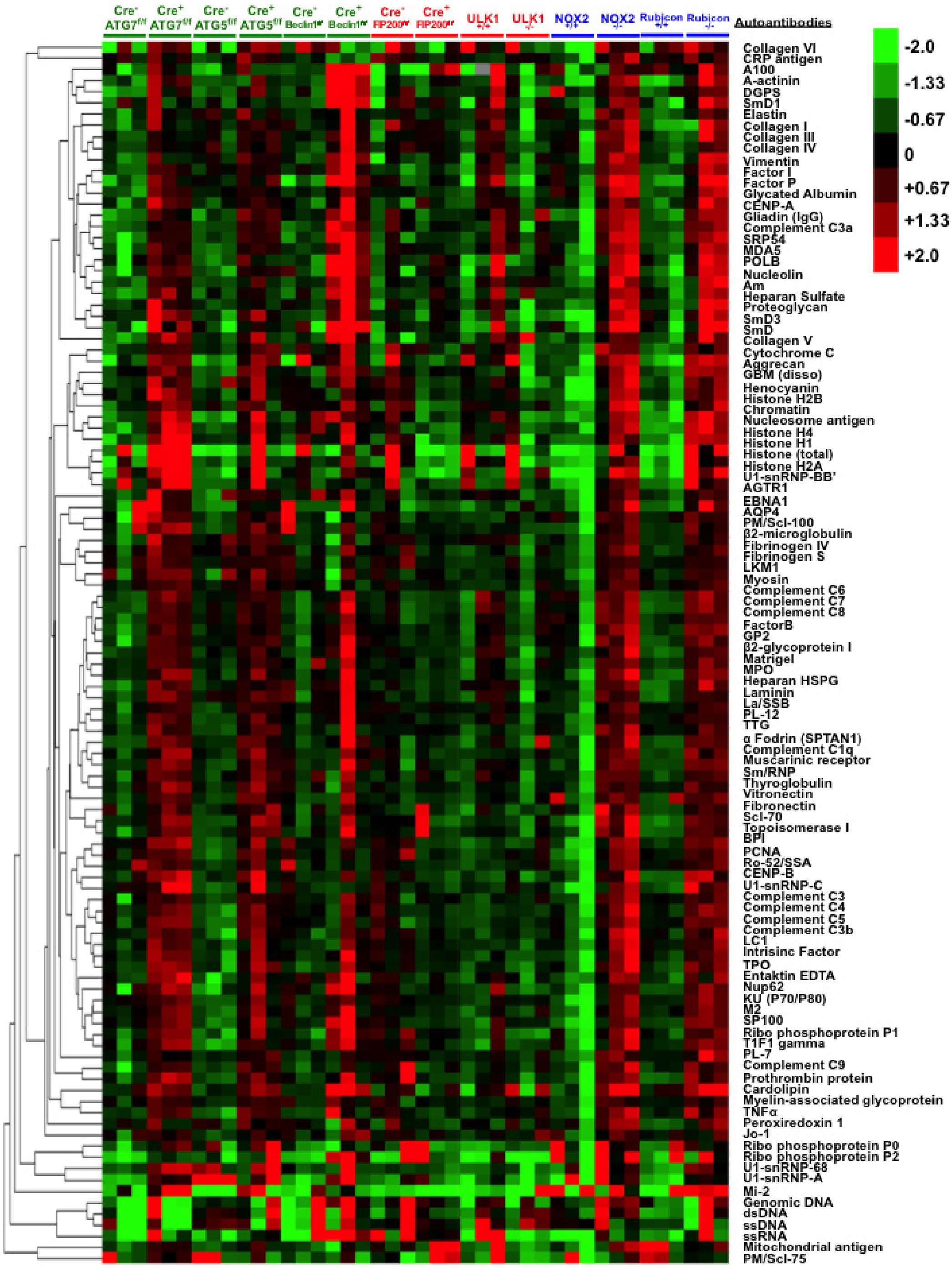
**Statistical analysis.** The statistical significance of differences in mean values was calculated using unpaired, two-tailed Student's *t*-test. *P* values less than 0.05 were considered statistically significant. No statistical methods were used to predetermine sample size. Experiments were not randomized, and the investigators were not blinded to allocation during experiments and outcome assessment.



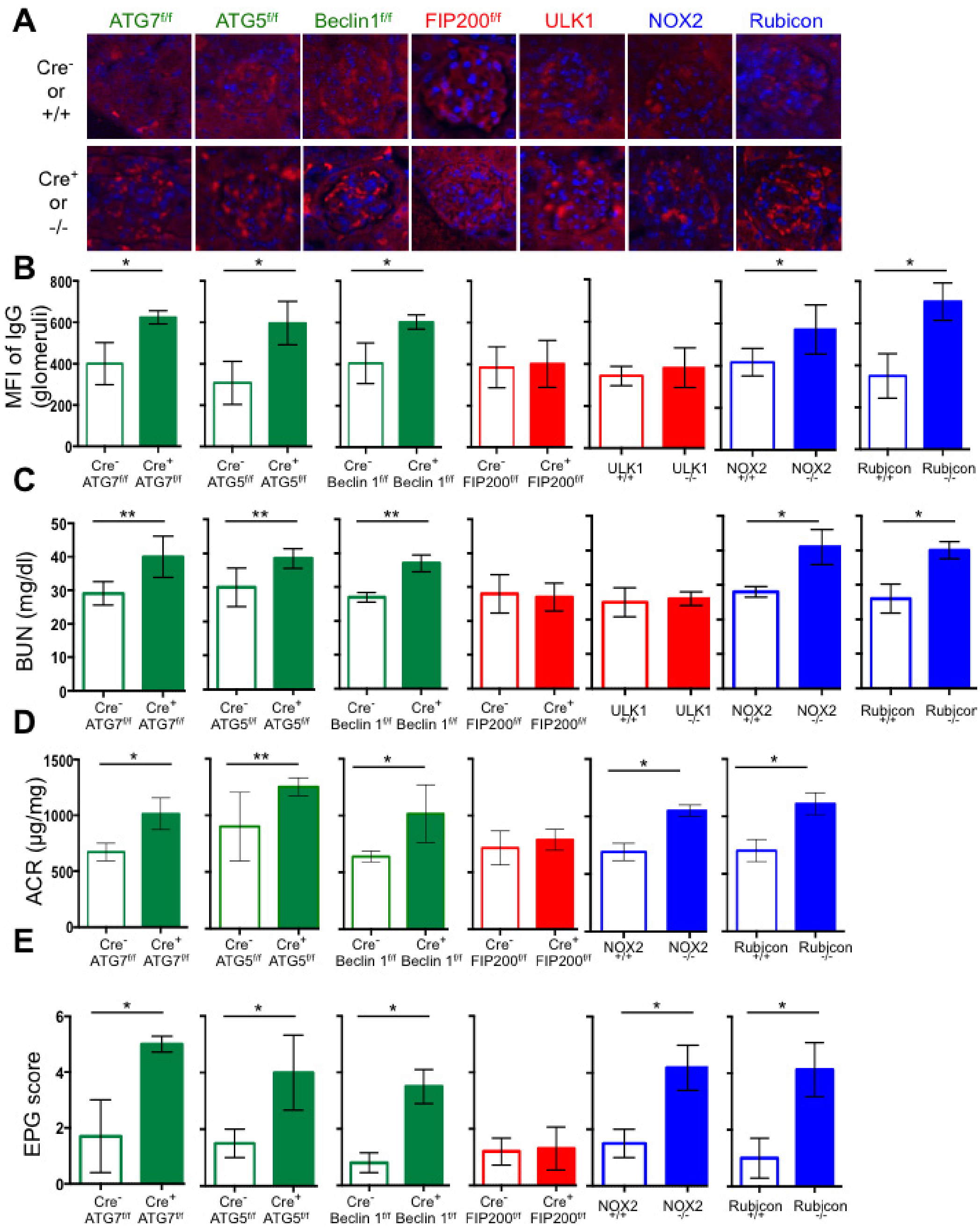

**Extended Data Figure 1 | LysM-Cre recombinase activity *in vivo*.** **a**, **b**, Bone marrow (**a**) and spleen (**b**) were obtained from wild-type and LysM-Cre$^{+}$ LSL–YFP reporter mice (R26-stop-EYFP) at 8 weeks of age and flow cytometry was performed to examine expression of YFP in the following cellular populations: macrophages (CD11b$^{+}$ F4/80$^{+}$), neutrophils (CD11b$^{+}$ Ly6G$^{+}$), monocytes (CD11b$^{+}$ CD115$^{+}$), conventional dendritic cells (CD11b$^{+}$ CD11c$^{+}$), plasmacytoid dendritic cells (CD11c$^{int}$ B220$^{+}$), CD8$\alpha^{+}$ dendritic cells (CD11c$^{+}$ CD8$\alpha^{+}$), eosinophils (CD11b$^{+}$ SiglecF$^{+}$), B cells (CD19$^{+}$ B220$^{+}$), CD4$^{+}$ T cells (CD3$^{+}$ CD4$^{+}$), and CD8$^{+}$ T cells (CD3$^{+}$ CD8$^{+}$). Error bars represent s.d. ($n = 4$).



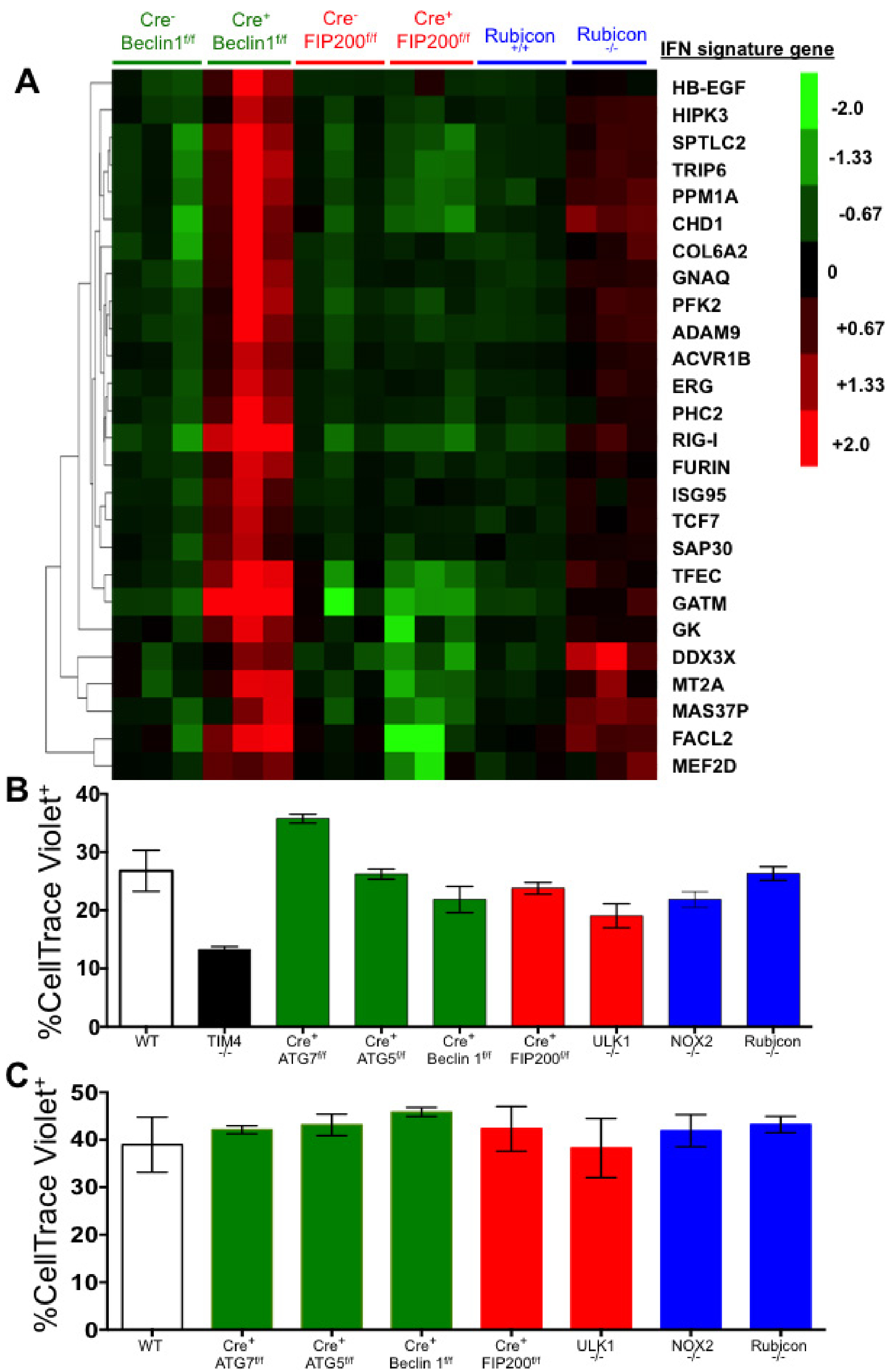

**Extended Data Figure 2 | Mice with LAP deficiencies display symptoms of immune activation. a**, Wild-type and deficient littermates were co-housed and aged for 52 weeks at SJCRH. Whole blood was collected at 52 weeks and analysed for differential blood count. Error bars represent s.d. LYM, lymphocytes; NEU, neutrophils; WBC, white blood cell s. **b**, **c**, Peripheral blood from *Rubcn*$^{+/+}$ and *Rubcn*$^{-/-}$ animals aged 52 weeks was analysed for immune cell populations. Neutrophils (singlets/CD3$^{-}$ CD19$^{-}$/Gr-1$^{hi}$ CD11b$^{+}$), monocytes (singlets/CD3$^{-}$ CD19$^{-}$/Gr-1$^{int}$ CD11b$^{+}$), activated T cells (singlets/CD3$^{+}$ CD4$^{+}$/CD44$^{+}$ CD62L$^{-}$ and singlets/CD3$^{+}$ CD8$^{+}$/CD44$^{+}$ CD62L$^{-}$), and central memory T cells (singlets/CD3$^{+}$ CD4$^{+}$/CD44$^{+}$ CD62L$^{+}$ and singlets/CD3$^{+}$ CD8$^{+}$/CD44$^{+}$ CD62L$^{+}$) were analysed and quantified. Error bars represent s.d. ($n = 5$, $^{**}P < 0.05$, Student's *t*-test). **d**, Spleens from wild-type and deficient littermates aged for 52 weeks were stained for anti-CD3 (top) or Ki67 (bottom) using immunohistochemistry. Representative images (original magnification, ×2.5) are shown ($n = 4$ per genotype). Error bars represent s.d. The colour scheme throughout represents LAP-deficient, autophagy-deficient genotypes (green), autophagy-deficient, LAP-sufficient (red), and autophagy-sufficient, LAP-deficient (blue).



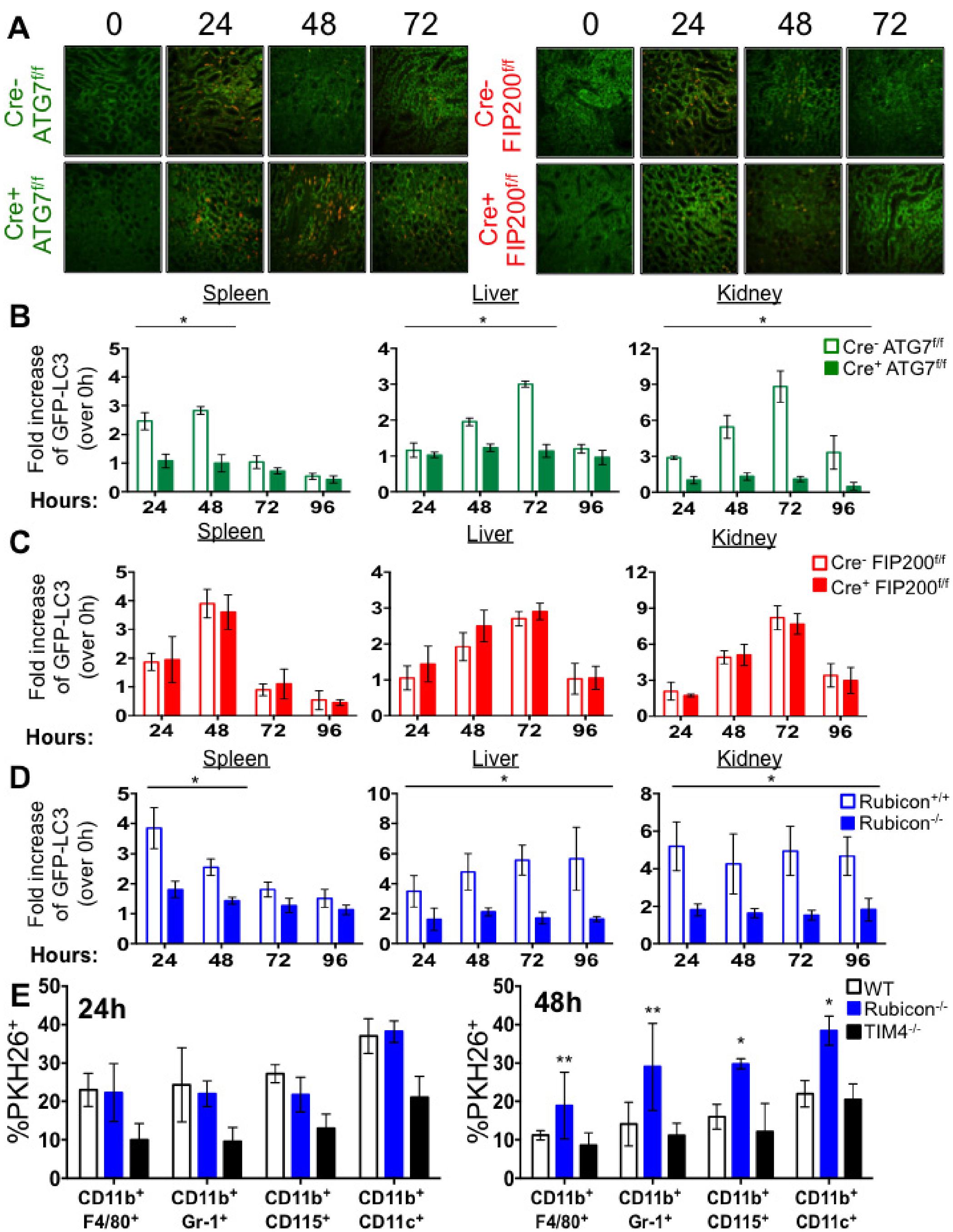

**Extended Data Figure 3 | Mice with LAP deficiencies display increased levels of circulating autoantibodies.** Serum from animals aged 52 weeks at SJCRH was analysed for autoantigens commonly associated with autoimmune and autoinflammatory disorders. The background subtracted signal intensity of each autoantigen was normalized to the average intensity of the total mouse IgG, which was included on the array as an internal control. IgG autoantibodies are shown, in triplicates per genotype.



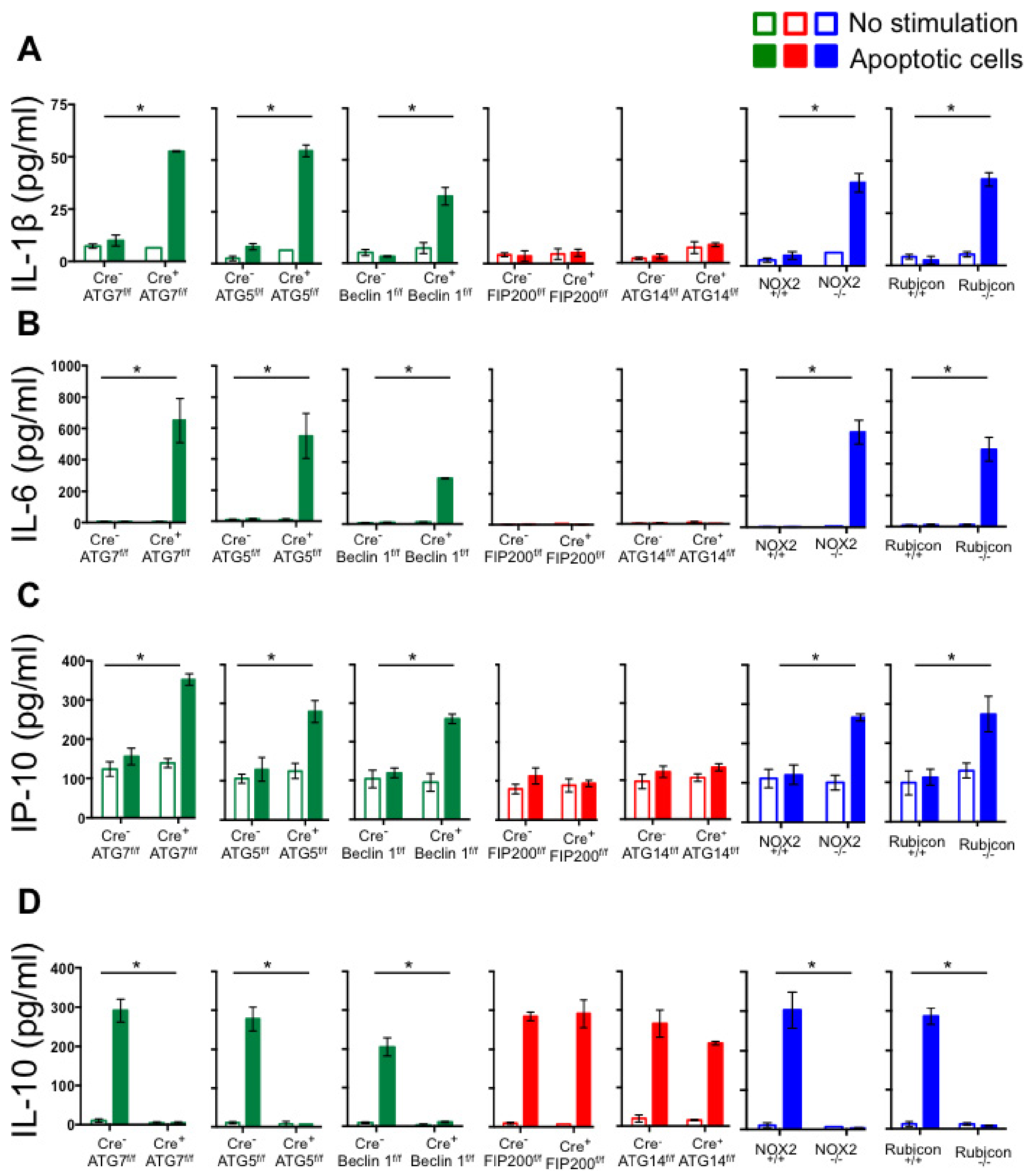

**Extended Data Figure 4 | Mice with LAP deficiencies display kidney pathology. a**, **b**, Wild-type and deficient littermates were co-housed and aged for 52 weeks at SJCRH. At 32 weeks, kidneys were obtained and stained for anti-IgG (red) and DAPI (blue) (**a**). Original magnification, ×100. MFI of anti-IgG staining in the glomeruli was calculated using Slidebook6 software (**b**). Error bars represent s.d. ($n > 15$ glomeruli per genotype, $*P < 0.001$, Student's $t$-test). **c**, At 52 weeks, serum was collected and analysed for blood urea nitrogen (BUN). **d**, At 52 weeks, urine was collected, and proteinuria was calculated as the ratio of albumin to creatinine (ACR). Error bars represent s.d. ($n > 4$ per genotype, $*P < 0.001$, $**P < 0.05$). **e**, At 52 weeks, kidneys were obtained and stained for haematoxylin and eosin. Kidneys were scored blindly for endocapillary proliferative glomerulonephritis (EPG) on a scale of 1 (no damage) to 5 (clear damage). For histological assessment, at least 24 glomeruli were evaluated for each genotype. Error bars represent s.d. ($*P < 0.001$, Student's $t$-test).



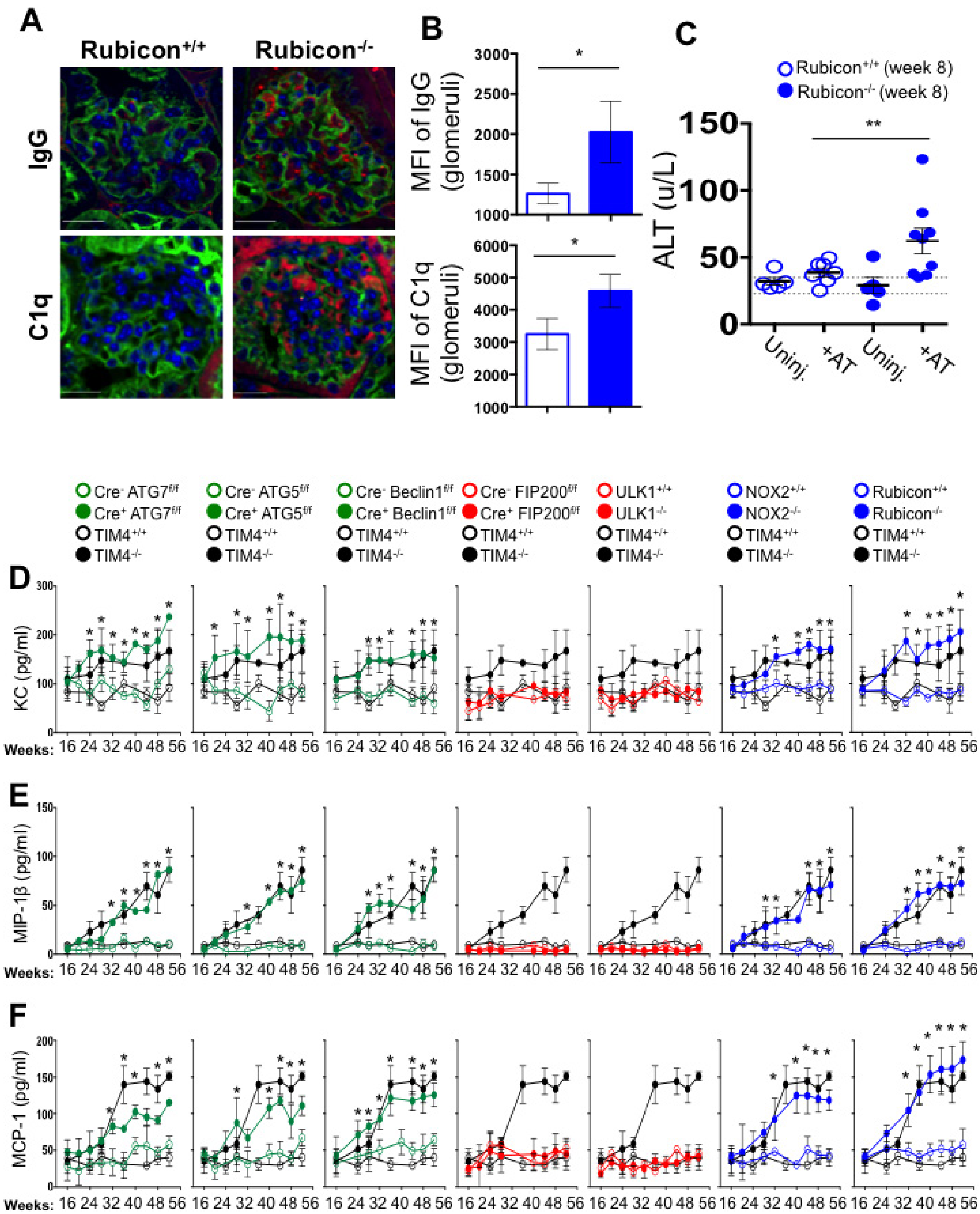

**Extended Data Figure 5 | Mice with LAP deficiencies display increased expression of the IFN signature but normal phagocytic capacity.** **a**, Wild-type and deficient littermates were co-housed and aged for 52 weeks at SJCRH. RNA was extracted from 52-week-old spleens and analysed for expression of genes associated with the IFN signature using Nanostring technology. Heatmap of Nanostring counts from the top 26 regulated genes in the IFN signature are shown in triplicate per genotype. **b**, UV-irradiated wild-type (WT) thymocytes were stained with CellTrace Violet and co-cultured (5:1) with bone-marrow-derived macrophages from wild-type and deficient genotypes for 45 min. Percentage phagocytosis (%CellTrace Violet$^+$) was quantified by flow cytometry (singlets/GFP$^+$ CellTrace Violet$^+$). **c**, Wild-type and deficient littermates were co-housed and aged for 52 weeks at SJCRH. Peritoneal macrophages were isolated after 3 days of intra-peritoneal injection of thioglycolate. UV-irradiated wild-type thymocytes were stained with CellTrace Violet and co-cultured (2:1) with peritoneal macrophages from wild-type and deficient genotypes for 1 h. Phagocytic efficiency (singlets/CellTrace Violet$^+$/F4/80$^+$) was quantified by flow cytometry (%CellTrace Violet$^+$). Error bars represent s.d. Data shown are representative of two independent experiments.



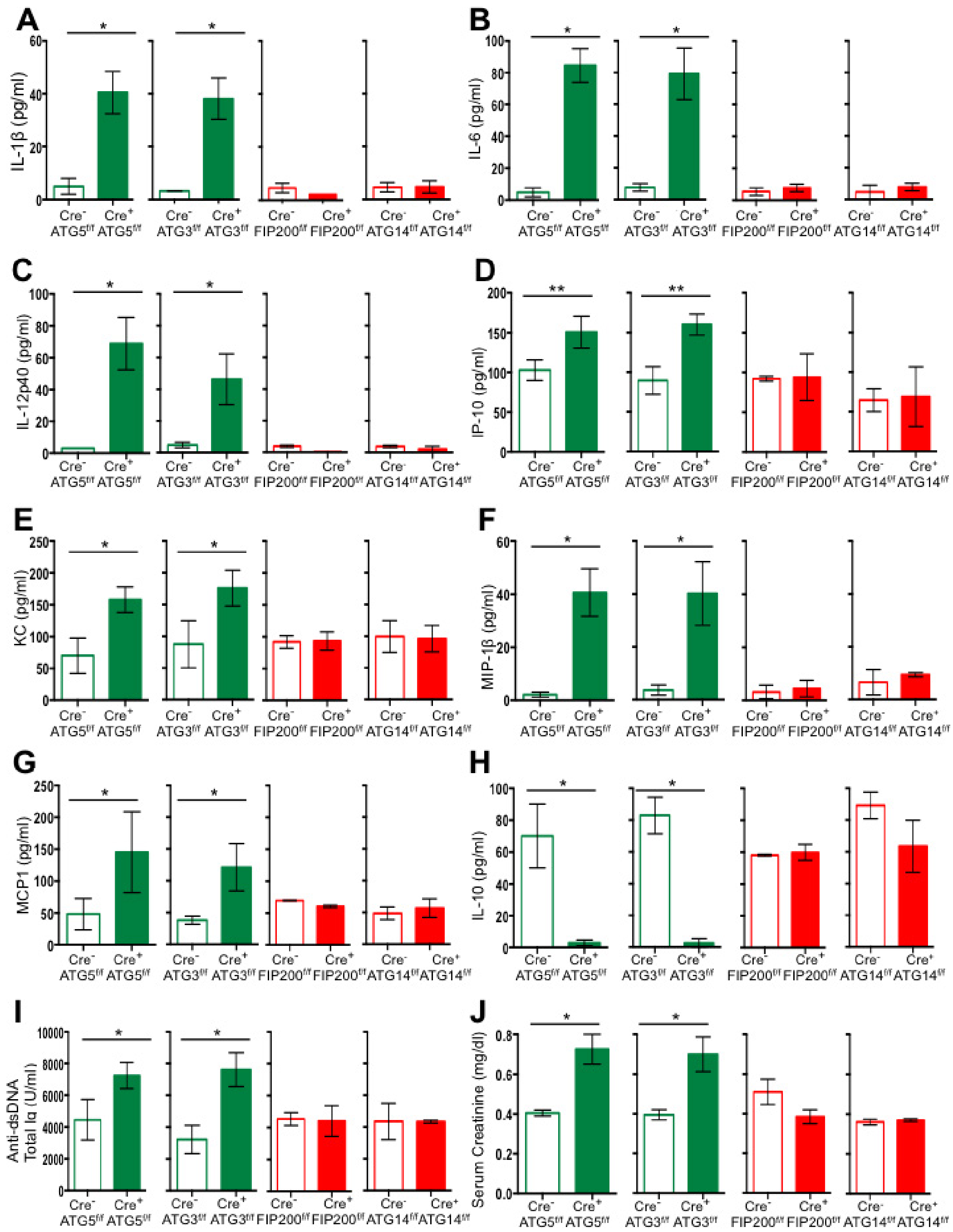

**Extended Data Figure 6 | Mice with LAP deficiencies display defective clearance of engulfed, dying cells. a**, $1 \times 10^7$ PKH26-labelled UV-irradiated wild-type thymocytes were injected intravenously into Cre$^-$ *Atg7*$^{f/f}$, Cre$^+$ *Atg7*$^{f/f}$, Cre$^-$ *Fip200*$^{f/f}$, or Cre$^+$ *Fip200*$^{f/f}$ animals (all GFP–LC3$^+$). Presence of labelled, apoptotic thymocytes was measured in kidney sections at 0, 24, 48, 72 and 96 h after transfer. Red cells are PKH26-labelled apoptotic thymocytes, and the kidney tissue is GFP–LC3. Representative images (original magnification, ×40) from two independent experiments are shown. **b**–**d**, Co-localization of lipidated GFP–LC3-II with engulfed dead cells was analysed by flow cytometry using digitonin treatment of spleen, liver and kidney of Cre$^-$ and Cre$^+$ *Atg7*$^{f/f}$ mice (**b**), Cre$^-$ and Cre$^+$ *Fip200*$^{f/f}$ mice (**c**), and *Rubcn*$^{+/+}$ and *Rubcn*$^{-/-}$ mice (**d**) at the indicated time points. **e**, $1 \times 10^7$ PKH26-labelled UV-irradiated wild-type thymocytes were injected intravenously into wild-type, *Rubcn*$^{-/-}$, or *Tim4*$^{-/-}$ animals. After 24 and 48 h, spleens were collected and stained with fluorescently conjugated surface markers for macrophages (CD11b$^+$ F4/80$^+$), neutrophils (CD11b$^+$ Gr-1$^+$), monocytes (CD11b$^+$ CD115$^+$), and dendritic cells (CD11b$^+$ CD11c$^+$). Phagocytic efficiency of each cell type (singlets/cell surface markers$^+$/PKH26$^+$) was quantified by flow cytometry (percentage PKH26). Data shown are representative of two independent experiments. Error bars represent s.d. (**$P < 0.05$, *$P < 0.001$, Student's *t*-test). The colour scheme represents LAP-deficient, autophagy-deficient genotypes (green), autophagy-deficient, LAP-sufficient (red), autophagy-sufficient, LAP-deficient (blue), and *Tim4*$^{+/+}$ and *Tim4*$^{-/-}$ (black).



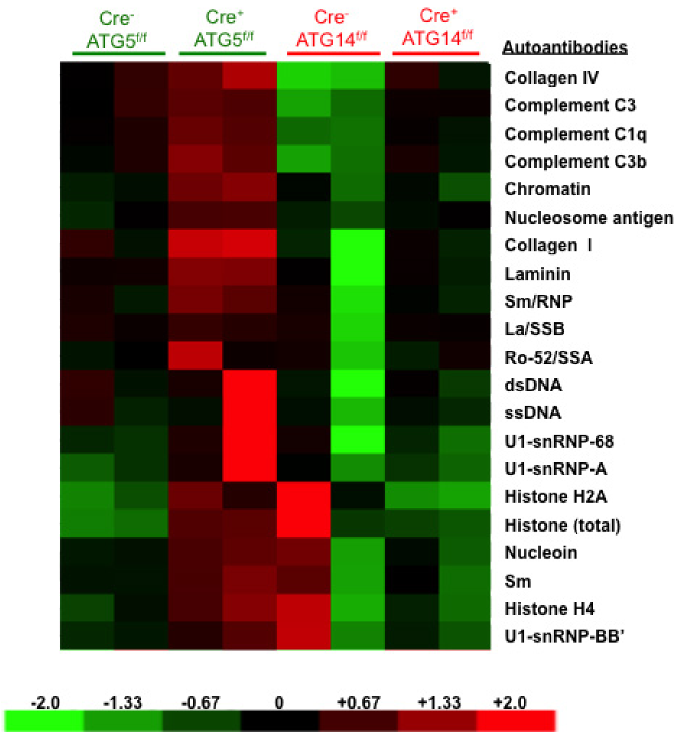

**Extended Data Figure 7 | LAP is required for the anti-inflammatory response to apoptotic cell engulfment *in vitro*. a–d**, UV-irradiated wild-type thymocytes were co-cultured with bone-marrow-derived macrophages from wild-type and deficient genotypes. Supernatant was collected at 24 h and analysed for IL-1β (**a**), IL-6 (**b**), IP-10 (**c**), and IL-10 (**d**) using Luminex technology. Error bars represent s.d. ($n = 4$, $^*P < 0.001$, Student's $t$-test). The colour scheme represents LAP-deficient, autophagy-deficient genotypes (green), autophagy-deficient, LAP-sufficient (red), autophagy-sufficient, and LAP-deficient (blue).




**Extended Data Figure 8 | Mice with LAP deficiencies display symptoms of an autoinflammatory disorder.** **a**–**c**, $2 \times 10^7$, UV-irradiated wild-type thymocytes were injected intravenously for 8 consecutive weeks into *Rubcn*$^{+/+}$ or *Rubcn*$^{-/-}$ animals (aged 6 weeks). After 8 weeks, kidneys were obtained and stained with DAPI (blue), wheat germ agglutinin (green), anti-IgG (red, top) and anti-C1q (red, bottom) (**a**). Original magnification, ×100. MFI of anti-IgG (top) and anti-C1q (bottom) staining in the glomeruli was calculated using Slidebook6 software (**b**). Error bars represent s.d. ($n > 15$ glomeruli per genotype, $*P < 0.001$, Student's *t*-test). After 8 weeks (week 8), serum was collected from uninjected and injected (+AT) animals (all 16 weeks of age) and analysed for alanine aminotransferase (ALT). Dots represent values from individual animals (**c**). Error bars represent s.e.m. ($**P < 0.05$, Student's *t*-test). **d**–**f**, Wild-type and deficient littermates were co-housed and aged for 52 weeks at SJCRH. Serum was collected every 4 weeks and analysed for KC (**d**), MIP-1β (**e**), and MCP1 (**f**) using Luminex technology. Error bars represent s.d. The colour scheme throughout represents LAP-deficient, autophagy-deficient genotypes (green), autophagy-deficient, LAP-sufficient (red), and autophagy-sufficient, LAP-deficient (blue). Values for one cohort of *Tim4*$^{+/+}$ and *Tim4*$^{-/-}$ animals are shown for comparison in all cases (black) in **a**–**c**.




**Extended Data Figure 9 | Mice with LAP deficiencies display symptoms of an autoinflammatory disorder.** Wild-type and deficient littermates were co-housed and aged for 52 weeks at Washington University. Serum was collected at 48–52 weeks and analysed for IL-1β (**a**), IL-6 (**b**), IL-12p40 (**c**), IP-10 (**d**), KC (**e**), MIP-1β (**f**), MCP1 (**g**), and IL-10 (**h**) using Luminex technology. Serum was analysed for anti-dsDNA antibodies (total Ig (**i**) and creatinine (**j**)). Error bars represent s.d. ($^{**}P < 0.001$, Student's *t*-test). The colour scheme throughout represents LAP-deficient, autophagy-deficient genotypes (green) and autophagy-deficient, LAP-sufficient (red).




**Extended Data Figure 10 | Mice with LAP deficiencies display increased levels of circulating autoantibodies.** Serum from animals aged 52 weeks at Washington University was analysed for autoantigens commonly associated with autoimmune and autoinflammatory disorders. IgG autoantibodies are shown, in duplicates per genotype.

doi:10.1038/nature17657

# Ubiquitination independent of E1 and E2 enzymes by bacterial effectors

Jiazhang Qiu[1], Michael J. Sheedlo[2], Kaiwen Yu[3], Yunhao Tan[1]†, Ernesto S. Nakayasu[4], Chittaranjan Das[2], Xiaoyun Liu[3] & Zhao-Qing Luo[1]

**Signalling by ubiquitination regulates virtually every cellular process in eukaryotes. Covalent attachment of ubiquitin to a substrate is catalysed by the E1, E2 and E3 three-enzyme cascade[1], which links the carboxy terminus of ubiquitin to the ε-amino group of, in most cases, a lysine of the substrate via an isopeptide bond. Given the essential roles of ubiquitination in the regulation of the immune system, it is not surprising that the ubiquitination network is a common target for diverse infectious agents[2]. For example, many bacterial pathogens exploit ubiquitin signalling using virulence factors that function as E3 ligases, deubiquitinases[3] or as enzymes that directly attack ubiquitin[4]. The bacterial pathogen *Legionella pneumophila* utilizes approximately 300 effectors that modulate diverse host processes to create a permissive niche for its replication in phagocytes[5]. Here we demonstrate that members of the SidE effector family of *L. pneumophila* ubiquitinate multiple Rab small GTPases associated with the endoplasmic reticulum. Moreover, we show that these proteins are capable of catalysing ubiquitination without the need for the E1 and E2 enzymes. A putative mono-ADP-ribosyltransferase motif critical for the ubiquitination activity is also essential for the role of the SidE family in intracellular bacterial replication in a protozoan host. The E1/E2-independent ubiquitination catalysed by these enzymes is energized by nicotinamide adenine dinucleotide, which activates ubiquitin by the formation of ADP-ribosylated ubiquitin. These results establish that ubiquitination can be catalysed by a single enzyme, the activity of which does not require ATP.**

The ability of the bacterial pathogen *L. pneumophila* to replicate within a phagocyte depends completely upon the Dot/Icm type IV secretion system that translocates hundreds of substrates (effectors) into host cells[6–8]. The activity of these effectors supports the biogenesis of the *Legionella*-containing vacuole (LCV), an area that is made permissive for bacterial replication by manipulating such diverse host processes as vesicle trafficking[5], protein translation[9], autophagy[10], cell migration[11], gene expression[12] and the biosynthesis of signalling lipids[13], often with sophisticated mechanisms[14]. With a few exceptions the roles of Dot/Icm effectors in *L. pneumophila* infection of its host are not fully understood because deletion of these genes individually often does not affect intracellular bacterial replication[5]. A biochemical function has been assigned to less than 10% of these effectors[5].

The SidE effector family contains four large proteins that are required for proficient intracellular bacterial replication[6,15]. PSI-BLAST analysis identified a putative mono-ADP-ribosyltransferase (mART) motif (R-S-ExE) in the central region of each of these proteins that is also present in such bacterial toxins as IotA[16], C3 exoenzyme[17] and ExoS[18] (Fig. 1a). Among these, the putative mART element in SdeA is $R_{766}$–$S_{820}$–$E_{860}S_{861}E_{862}$, a catalytic motif found in enzymes that transfer the ADP-ribosyl group from nicotinamide adenine dinucleotide (NAD) to arginine residues[19]. To examine its role in SdeA-mediated yeast toxicity[20,21], we created the $SdeA_{E/A}$ mutant, in which Glu860 and Glu862 were mutated to alanine. This mutant has completely lost its toxicity to yeast and was also defective in inhibiting the secretion of the secreted form of the embryonic alkaline phosphatase (SEAP)[22] by mammalian cells (Fig. 1b, c). SidE, SdeB and SdeC also significantly inhibited SEAP secretion in a manner dependent upon the predicted mART motif (Extended Data Fig. 1a). These results suggest that the putative mART motif is essential for the activity of the SidE family effectors.

A mutant missing the SidE family (Δ*sidE*) shows attenuated virulence against the protozoan host *Dictyostelium discoideum*[15] (Fig. 2a). Expression of wild-type SdeA but not the $SdeA_{E/A}$ mutant in a Δ*sidE* strain almost completely restored its ability to grow within the host (Fig. 2a, b). In *D. discoideum*, LCVs containing wild-type bacteria efficiently recruit endoplasmic reticulum (ER) markers such as the GFP–HDEL fusion to their surface, which is a hallmark of *L. pneumophila* infection[23,24]. Similar to its defects in intracellular growth, the Δ*sidE* mutant no longer recruited GFP–HDEL to its vacuoles, even at 10 h post infection (Fig. 2c, d and Extended Data Fig. 1b, c). Again, SdeA but not $SdeA_{E/A}$ complemented such defects (Fig. 2c, d). Thus, the putative mART motif is important for the function of the SidEs during bacterial infection.

Next we attempted to determine the potential ADP-ribosyltransferase activity of SdeA. Despite extensive efforts, we were unable to detect SdeA-mediated ADP-ribosylation of eukaryotic proteins (Extended Data Fig. 2a), suggesting that this protein possesses a different biochemical activity. During *L. pneumophila* infection, members of the SidE family are transiently associated with the LCV[15], an organelle resembling the ER[23]. Because Rab small GTPases are a common target of *L. pneumophila* effectors[25], we examined whether SdeA attacks any of the ER-associated Rab proteins[26] by co-expressing 4×Flag-tagged Rab1, Rab6A, Rab30 or Rab33b with this effector in mammalian cells. A clear shift in molecular mass was observed for all four Rab proteins purified from cells co-transfected with SdeA but not $SdeA_{E/A}$ (Fig. 3a, left and middle panels). Such a molecular mass shift did not occur for the endosomal Rab5 or the cytoskeletal small GTPase Rac1 (Fig. 3a, right panel), indicating potential substrate specificity. Among the proteins potentially modified by SdeA, the modification of Rab33b was the most extensive, suggesting that this protein is a preferred substrate. The molecular mass shift in Rab33b also was observed when it was co-expressed with other members of the SidE family (Extended Data Fig. 2b). To determine whether the potential post-translational modification occurs during bacterial infection, we infected mammalian cells expressing 4×Flag-Rab33b with *L. pneumophila*. Rab33b of higher molecular mass was detected in samples infected with the wild-type strain but not with strains lacking the Dot/Icm transporter or the SidE family (Fig. 3b). The defect in Rab33b modification exhibited by the Δ*sidE* strain can be complemented by expressing SdeA but not $SdeA_{E/A}$ (Fig. 3b). A similar SidE-dependent molecular mass

[1]Purdue Institute for Inflammation, Immunology and Infectious Disease and Department of Biological Sciences, Purdue University, West Lafayette, Indiana 47907, USA. [2]Department of Chemistry, Purdue University, 560 Oval Drive, West Lafayette, Indiana 47907, USA. [3]Institute of Analytical Chemistry and Synthetic and Functional Biomolecules Center, College of Chemistry and Molecular Engineering, Peking University, Beijing 100871, China. [4]Biological Science Division, Pacific Northwest National Laboratory, Richland, Washington 99352, USA. †Present address: Division of Gastroenterology, Boston Children's Hospital, Harvard Medical School, Boston, Massachusetts 02115, USA.

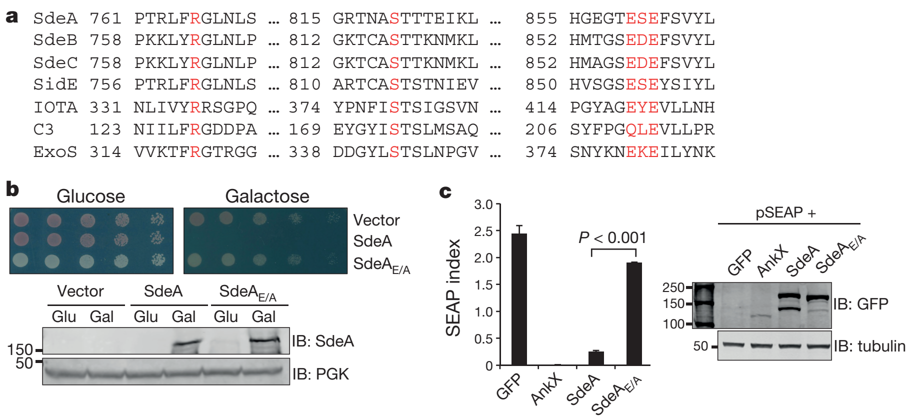



**Figure 1 | A putative mono-ADP-ribosyltransferase (mART) motif is important for yeast toxicity of SdeA. a**, Alignment of the central region of the SidE family members and several toxins with mART activity. Proteins identified by PSI-BLAST were manually aligned. Shown mART toxins are IotA from *Clostridium perfringens*[16], the C3 exoenzyme from *Clostridium botulinum*[17] and ExoS from *Pseudomonas aeruginosa*[18]. Residues important for the mART motif were highlighted in red. **b**, **c**, The mART is essential for yeast toxicity and for secretion inhibition by SdeA. Yeast cells were spotted on the indicated medium for 3 days before image acquisition. The secretion of SEAP was examined in 293T cells transfected to express SEAP and GFP-tagged testing proteins; the strong SEAP inhibitor AnkX[22] was used as a control. Error bars represent s.e.m. ($n = 3$). The expression of the proteins (the lower panel in **b** for yeast and the right panel in **c** for mammalian cells) was probed with indicated antibodies. The PGK (3-phosphoglyceric phosphokinase) and tubulin were probed as a loading control, respectively. SdeA$_{E/A}$, SdeA with Glu860 and Glu862 mutated to Ala. IB, immunoblotting. The yeast toxicity results in **b** and protein levels in **b** and **c** are from one representative of three independent experiments. The SEAP results in **c** are one representative done in triplicate from three independent experiments. **b**, **c**, Uncropped blots are shown in Supplementary Fig. 1.

shift also occurred to Rab1 during bacterial infection (Extended Data Fig. 2c). Thus, SdeA induces a biochemical modification of multiple ER-associated Rabs, and at least Rab33b and Rab1 are substrates during bacterial infection.

We next determined the nature of the SdeA-induced post-translational modification by mass spectrometric analysis of 4×Flag–Rab33b purified from 293T cells expressing SdeA. Ubiquitin fragments were only detected in Rab33b of higher molecular mass (Fig. 3c, d and Extended Data Fig. 3a). Similar results were obtained in Rab33b from cells infected with wild-type *L. pneumophila* (Fig. 3e, f). These results suggest that Rab33b is involved in the formation of the LCV and that SdeA induces ubiquitination of Rab33b in a process that requires the putative mART motif. Indeed, overexpression of wild type Rab33b but not its dominant negative or dominant positive mutants[27], inhibits the formation of vacuoles containing large number (>10) of bacteria (Fig. 3g and Extended Data Fig. 3b).

Ubiquitination requires enzymes E1, E2 and E3 which activates, conjugates and transfers the ubiquitin molecule to the substrate,

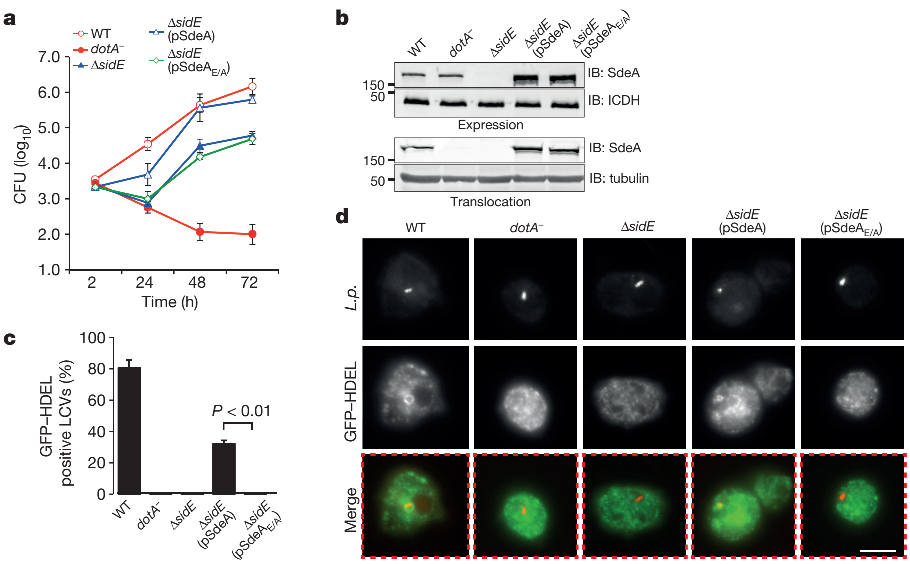



**Figure 2 | The predicted mART motif is essential for the role of SdeA in intracellular bacterial growth. a**, The indicated bacterial strains were used to infect *D. discoideum* and the bacterial yields were monitored at 24-h intervals. Note that SdeA but not the SdeA$_{E/A}$ mutant restored the defect exhibited by the Δ*sidE* strain. CFU, colony-forming units. **b**, Expression and Dot/Icm-mediated translocation of SdeA and SdeA$_{E/A}$. The bacteria used for infections were probed for protein expression; the metabolic enzyme isocitrate dehydrogenase (ICDH) was probed as a loading control (top panel). Saponin-soluble fractions of infected cells were probed for translocated SdeA with tubulin as a loading control (bottom panel). **c**, **d**, *L. pneumophila* was used to infect a strain of *D. discoideum* stably expressing the ER retention fusion GFP–HDEL and the recruitment of the ER marker to the phagosome was evaluated 2 h after infection. IB, immunoblotting. Results in **a** and **c** are from one representative experiment done in triplicate from three independent experiments; error bars represent s.e.m. ($n = 3$). Results in **b** and **d** are one representative from three independent experiments. Scale bar, 5 μm. **b**, Uncropped blots are shown in Supplementary Fig. 1.

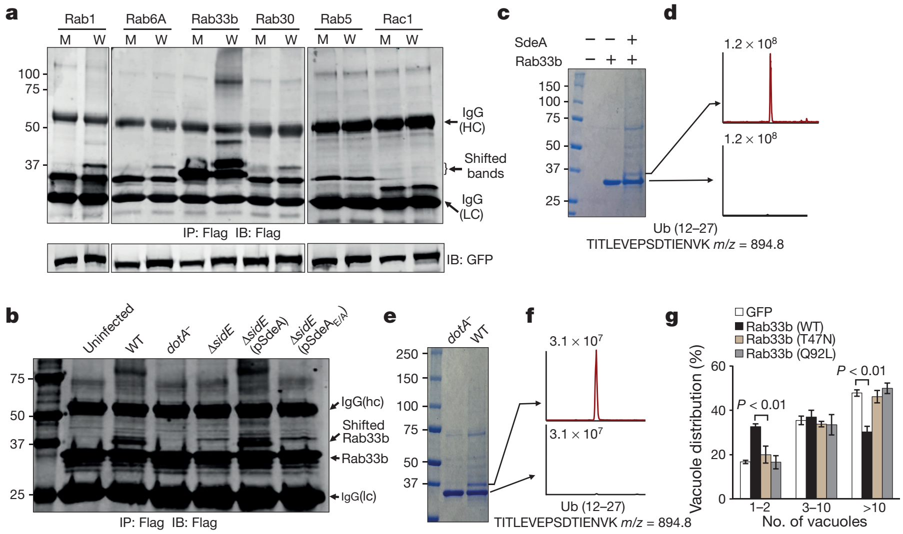



**Figure 3 | SdeA induces a posttranslational modification on multiple ER-associated Rab proteins. a**, Lysates of 293T cells co-transfected to express SdeA and Flag-tagged small GTPases were subjected to immunoprecipitation with Flag beads and the products were probed with the Flag-specific antibody. Note the appearance of shifted bands for Flag-tagged ER-associated Rabs but not for Rab5 and Rac1. M, SdeA$_{E/A}$; W, SdeA; IgG (HC) and IgG (LC) indicate IgG heavy and light chains, respectively. **b**, SdeA-dependent post-translational modification of Rab33b during bacterial infection. Cells expressing Flag–Rab33b were infected with relevant *L. pneumophila* strains for 2 h and Flag–Rab33b purified from cell lysates was probed by immunoblotting. **c**–**f**, SdeA induces Rab33b ubiquitination. Flag–Rab33b purified from cells co-expressing SdeA (**c**) or infected with wild-type *L. pneumophila* (**e**) was subjected to mass spectrometric analysis and tryptic ubiquitin fragments were identified in proteins of the shifted bands (**d**, **f**). **g**, Overexpression of Rab33b restricts intracellular bacterial growth. COS1 cells transfected with Rab33b and the indicated mutants were infected with *L. pneumophila* and the formation of replicative vacuoles was determined. IB, immunoblotting. Data shown are one representative experiment of three independent experiments (**a**–**f**); results in **g** are one representative done in triplicate from three independent experiments. Error bars represent s.e.m. ($n = 3$). **a**–**c**, **e**, Uncropped blots and gel images are shown in Supplementary Fig. 1.

respectively[1]. We thus used *in vitro* reactions to determine whether SdeA directly participates in the ubiquitination of Rab33b. In a series of reactions each containing E1 and one of several E2 enzymes, no ubiquitination of Rab33b was detected (Extended Data Fig. 3c). We thus tested the hypothesis that an unknown E2 is required for the activity of SdeA by adding cell lysates to the reactions, which led to ubiquitination of Rab33b in an mART-dependent manner (Fig. 4a). Unexpectedly, ubiquitination still occurred in reactions receiving heat-treated cell lysates (Fig. 4a, lane 3), suggesting that both E1 and the putative SdeA-specific E2 are heat-stable or that SdeA is able to catalyse ubiquitination by itself but only in the presence of heat-stable molecule(s) from cells. To distinguish between these two possibilities, we added *E. coli* lysates to the reaction. Notably, ubiquitination of Rab33b did occur (Fig. 4a, lane 4). These results demonstrate that SdeA catalyses E1/E2-independent ubiquitination in a process that requires one or more heat-stable molecules present in cells.

Classic ubiquitination requires the conserved E1 that activates ubiquitin in a process powered by hydrolysis of ATP, which binds the enzyme in a $Mg^{2+}$-dependent manner[1]. We thus determined the requirement of these molecules in SdeA-mediated ubiquitination. Because of the importance of the mART motif in the cleavage of NAD by canonical ADP-ribosyltransferases[19], we included this compound in our reactions. In reactions containing NAD, $Mg^{2+}$ and ATP, ubiquitination of Rab33b occurred (Fig. 4b, lane 2). Yet, when NAD was withdrawn, no ubiquitination was detected (Fig. 4b, lane 3). In line with this observation, ubiquitination occurred in reactions containing NAD but not ATP or $Mg^{2+}$ (Fig. 4b, lanes 4 and 5). Heat-treated NAD is active, which is consistent with the fact that boiled cell lysates allowed SdeA to function (Fig. 4b, lane 8). Exogenous NAD is sufficient for the activity of SdeA that had been dialysed against a buffer containing EDTA (Extended Data Fig. 4a), suggesting that this compound is the only co-factor required for the activity. SdeA$_{E/A}$ is unable to catalyse the modification even in the presence of NAD (Fig. 4b, lane 9). Under this condition, both Rab1 and Rab6A were ubiquitinated by SdeA (Extended Data Fig. 4b). Similarly, SidE, SdeB and SdeC ubiquitinated Rab33b (Extended Data Fig. 4c). Consistently, SdeA does not detectably ADP-ribosylate Rab33b or Rab1 (Extended Data Fig. 5a).

Since ubiquitin ligases often self-modify[1], we incubated SdeA with glutathione *S*-transferase (GST)-tagged ubiquitin to probe such self-ubiquitination. Proteins of higher molecular mass were detected in reactions containing SdeA but not SdeA$_{E/A}$, again in a NAD-dependent manner (Fig. 4c). The central domain of SdeA remains toxic to yeast[20], suggesting that it is still biochemically active. Indeed, SdeA$_{178\text{-}1000}$ robustly ubiquitinates itself and Rab33b in a manner that requires both NAD and the mART motif (Fig. 4d). These results demonstrate that the N-terminal deubiquitinase (DUB) domain[28] of SdeA does not interfere with its ubiquitin conjugation activity. Indeed, the SdeA$_{C118A}$ mutant defective in the DUB activity[28] catalyses ubiquitination indistinguishably to that of the wild-type protein (Extended Data Fig. 5b, c).

Mass spectrometric and mutational analyses revealed that Arg42 of ubiquitin is important for SdeA-mediated, but not for canonical ubiquitination catalysed by the E1-E2-E3 cascade (Extended Data Fig. 6a, b). Consistent with these results, SdeA ubiquitinates Rab33b with all lysine variants of ubiquitin, as well as the ubiquitin derivative containing an alanine substitution in the last two glycine residues or with six histidine residues attached to its carboxy terminus (Extended Data Fig. 6c–e). Further, ubiquitination catalysed by SdeA is insensitive to the cysteine alkylation agent maleimide, suggesting that a cysteine conjugation of ubiquitin does not form during the reaction (Extended Data Fig. 7). Finally, ubiquitination by SdeA affected the GTP loading and hydrolysis activity of Rab33b but did not detectably affect its stability (Fig. 3a and Extended Data Fig. 8). The nucleotide binding status of Rab33b did not affect its suitability as the substrate of SdeA (Extended Data Fig. 8e).

We detected AMP, nicotinamide, ubiquitin and NAD in SdeA-catalysed reactions (Extended Data Fig. 9). The release of AMP suggests the formation of an ubiquitin-AMP adduct during the reaction. Yet, the ubiquitin-AMP adduct could not be detected by $^{32}$P-$\alpha$-NAD or by TCA precipitation followed by high-performance liquid chromatography–mass spectrometry (HPLC–MS) (Extended Data Fig. 10a). The release of nicotinamide and the requirement of Arg42 of ubiquitin

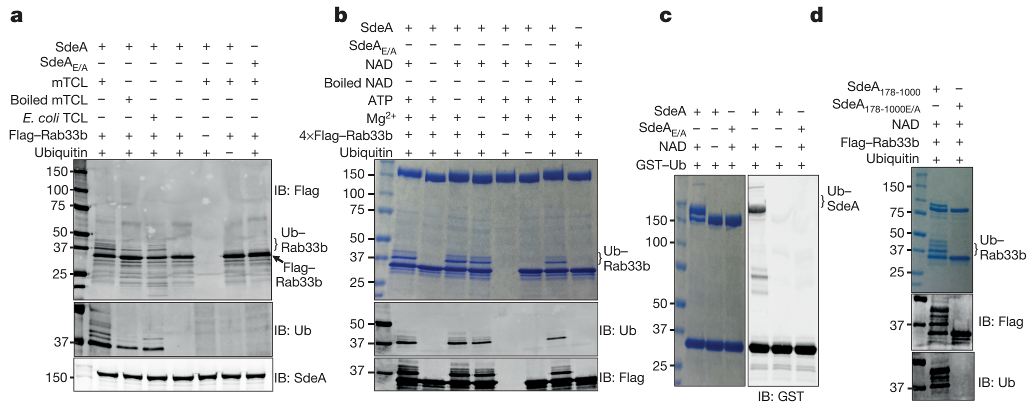


**Figure 4 | SdeA catalyses ubiquitination independent of E1 and E2.** **a**, A heat-stable molecule from cells is required for ubiquitination induced by SdeA. Reactions resolved by SDS–PAGE were probed with the indicated antibodies. Note the production of ubiquitinated Rab33b in reactions containing boiled mammalian (m) cell lysates and *E. coli* lysates. TCL, total cell lysates. **b**, NAD is required for SdeA-catalysed ubiquitination. Ubiquitinated Rab33b and SdeA were probed by Coomassie staining or by immunoblotting (IB) with antibodies specific for ubiquitin or Flag. **c**, Self-ubiquitination by SdeA. SdeA or $SdeA_{E/A}$ was incubated with GST–ubiquitin and NAD; ubiquitination was detected by immunoblotting or by Coomassie staining. Note the formation of the high molecular mass self-ubiquitinated SdeA when GST–ubiquitin was included in the reactions. **d**, Ubiquitination catalysed by the central domain of SdeA. $SdeA_{178-1000}$ or $SdeA_{178-1000E/A}$ was used for ubiquitination of Rab33b and the products were probed by Coomassie staining or by immunoblotting. **a**–**d**, Similar results were obtained from four experiments. Uncropped blots are shown in Supplementary Fig. 1.

implied ADP-ribosylation of this side chain as a possible step before ubiquitin conjugation, which is consistent with the requirement of the R-S-ExE motif found in members of the SidE protein family. Thus, we probed the reaction intermediate by obtaining $SdeA_{519-1100}$, a fragment that retained the ability to modify Rab33b but had lost the self-ubiquitination activity (Extended Data Fig. 10b, c). Incubation of $SdeA_{519-1100}$ with NAD and ubiquitin led to the release of nicotinamide (Extended Data Fig. 10d), suggesting the formation of ADP-ribosylated ubiquitin. Furthermore, inclusion of $^{32}$P-$\alpha$-NAD in the reaction produced $^{32}$P-labelled ubiquitin in an Arg42-dependent manner and the ADP-ribosyl moiety linked to Arg42 of ubiquitin can be detected by mass spectrometric analysis (Extended Data Fig. 10e–g). Thus, ADP-ribosylated ubiquitin is the reaction intermediate. The production of AMP in reactions with full-length SdeA could be a subsequent step in the attack of an acceptor nucleophile (from the Rab proteins or SdeA itself in the self-conjugation reaction) on the ADP-ribosylated ubiquitin leading to the modification of the target protein.

In a canonical ubiquitination reaction, ubiquitin activated by E1 is delivered to E2 to form the E2~Ub thioester. For the E3 ligases of the RING family, ubiquitin is directly transferred from the E2 to a substrate facilitated by the ligases, whereas members of the HECT and RBR E3 families transfer ubiquitin to a catalytic cysteine in the E3 before delivering it to the substrate[1]. Clearly, SdeA defines an all-in-one ubiquitin conjugation enzyme that directly activates ubiquitin; the fact that $SdeA_{519-1100}$ defective in auto-ubiquitination can still modify Rab33b suggests that the activated ubiquitin is directly transferred to the substrate.

The discovery that ubiquitin can be modified by ADP-ribosylation expands the post-translational modification on this prevalent signalling molecule, which has been shown to be modified by acetylation and phosphorylation[29]. Whether ADPR~Ub itself is directly used to modify proteins is unknown, but it is clear that such modifications can potentially lead to significant expansion of the ubiquitin code and its functions in cellular processes and disease development[29]. The mART motif is present in a family of mammalian proteins, some of which are unable to catalyse ADP-ribosylation[30]. In light of the mART-dependent ubiquitination activity of SdeA, it will be interesting to determine whether any of these mART-containing proteins is capable of catalysing ubiquitination, and if so, whether the reaction requires E1 and E2. The identification of eukaryotic mART proteins with such a capability will surely expand the spectrum of cellular processes regulated by ubiquitination.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 2 October 2015; accepted 14 March 2016.**
**Published online 6 April; corrected online 4 May 2016**
**(see full-text HTML version for details).**

1. Komander, D. & Rape, M. The ubiquitin code. *Annu. Rev. Biochem.* **81,** 203–229 (2012).
2. Zinngrebe, J., Montinaro, A., Peltzer, N. & Walczak, H. Ubiquitin in the immune system. *EMBO Rep.* **15,** 28–45 (2014).
3. Zhou, Y. & Zhu, Y. Diversity of bacterial manipulation of the host ubiquitin pathways. *Cell. Microbiol.* **17,** 26–34 (2015).
4. Cui, J. *et al.* Glutamine deamidation and dysfunction of ubiquitin/NEDD8 induced by a bacterial effector family. *Science* **329,** 1215–1218 (2010).
5. Xu, L. & Luo, Z. Q. Cell biology of infection by *Legionella pneumophila*. *Microbes Infect.* **15,** 157–167 (2013).
6. Luo, Z. Q. & Isberg, R. R. Multiple substrates of the *Legionella pneumophila* Dot/Icm system identified by interbacterial protein transfer. *Proc. Natl Acad. Sci. USA* **101,** 841–846 (2004).
7. Huang, L. *et al.* The E Block motif is associated with *Legionella pneumophila* translocated substrates. *Cell. Microbiol.* **13,** 227–245 (2011).
8. Lifshitz, Z. *et al.* Computational modeling and experimental validation of the Legionella and Coxiella virulence-related type-IVB secretion signal. *Proc. Natl Acad. Sci. USA* **110,** E707–E715 (2013).
9. Fontana, M. F. *et al.* Secreted bacterial effectors that inhibit host protein synthesis are critical for induction of the innate immune response to virulent *Legionella pneumophila*. *PLoS Pathog.* **7,** e1001289 (2011).
10. Choy, A. *et al.* The *Legionella* effector RavZ inhibits host autophagy through irreversible Atg8 deconjugation. *Science* **338,** 1072–1076 (2012).
11. Simon, S., Wagner, M. A., Rothmeier, E., Muller-Taubenberger, A. & Hilbi, H. Icm/Dot-dependent inhibition of phagocyte migration by *Legionella* is antagonized by a translocated Ran GTPase activator. *Cell. Microbiol.* **16,** 977–992 (2014).
12. Rolando, M. *et al. Legionella pneumophila* effector RomA uniquely modifies host chromatin to repress gene expression and promote intracellular bacterial replication. *Cell Host Microbe* **13,** 395–405 (2013).
13. Hsu, F. *et al.* Structural basis for substrate recognition by a unique *Legionella* phosphoinositide phosphatase. *Proc. Natl Acad. Sci. USA* **109,** 13567–13572 (2012).
14. Zhu, W. & Luo, Z. Q. Cell biology and immunology lessons taught by *Legionella pneumophila*. *Sci. China Life Sci.* **59,** 3–10 (2016).
15. Bardill, J. P., Miller, J. L. & Vogel, J. P. IcmS-dependent translocation of SdeA into macrophages by the *Legionella pneumophila* type IV secretion system. *Mol. Microbiol.* **56,** 90–103 (2005).
16. Sakurai, J., Nagahama, M., Oda, M., Tsuge, H. & Kobayashi, K. *Clostridium perfringens* iota-toxin: structure and function. *Toxins (Basel)* **1,** 208–228 (2009).
17. Wilde, C. & Aktories, K. The Rho-ADP-ribosylating C3 exoenzyme from *Clostridium botulinum* and related C3-like transferases. *Toxicon* **39,** 1647–1660 (2001).

18. Ganesan, A. K., Frank, D. W., Misra, R. P., Schmidt, G. & Barbieri, J. T. *Pseudomonas aeruginosa* exoenzyme S ADP-ribosylates Ras at multiple sites. *J. Biol. Chem.* **273,** 7332–7337 (1998).
19. Simon, N. C., Aktories, K. & Barbieri, J. T. Novel bacterial ADP-ribosylating toxins: structure and function. *Nature Rev. Microbiol.* **12,** 599–611 (2014).
20. Havey, J. C. & Roy, C. R. Toxicity and SidJ-mediated suppression of toxicity require distinct regions in the SidE family of *Legionella pneumophila* effectors. *Infect. Immun.* **83,** 3506–3514 (2015).
21. Jeong, K. C., Sexton, J. A. & Vogel, J. P. Spatiotemporal regulation of a *Legionella pneumophila* T4SS substrate by the metaeffector SidJ. *PLoS Pathog.* **11,** e1004695 (2015).
22. Tan, Y., Arnold, R. J. & Luo, Z. Q. *Legionella pneumophila* regulates the small GTPase Rab1 activity by reversible phosphorylcholination. *Proc. Natl Acad. Sci. USA* **108,** 21212–21217 (2011).
23. Swanson, M. S. & Isberg, R. R. Association of *Legionella pneumophila* with the macrophage endoplasmic reticulum. *Infect. Immun.* **63,** 3609–3620 (1995).
24. Liu, Y. & Luo, Z. Q. The *Legionella pneumophila* effector SidJ is required for efficient recruitment of endoplasmic reticulum proteins to the bacterial phagosome. *Infect. Immun.* **75,** 592–603 (2007).
25. Sherwood, R. K. & Roy, C. R. A. Rab-centric perspective of bacterial pathogen-occupied vacuoles. *Cell Host Microbe* **14,** 256–268 (2013).
26. Ortiz Sandoval, C. & Simmen, T. Rab proteins of the endoplasmic reticulum: functions and interactors. *Biochem. Soc. Trans.* **40,** 1426–1432 (2012).
27. Itoh, T. *et al.* Golgi-resident small GTPase Rab33B interacts with Atg16L and modulates autophagosome formation. *Mol. Biol. Cell* **19,** 2916–2925 (2008).
28. Sheedlo, M. J. *et al.* Structural basis of substrate recognition by a bacterial deubiquitinase important for dynamics of phagosome ubiquitination. *Proc. Natl Acad. Sci. USA* **112,** 15090–15095 (2015).
29. Herhaus, L. & Dikic, I. Expanding the ubiquitin code through post-translational modification. *EMBO Rep.* **16,** 1071–1083 (2015).
30. Glowacki, G. *et al.* The family of toxin-related ecto-ADP-ribosyltransferases in humans and the mouse. *Protein Sci.* **11,** 1657–1670 (2002).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** We thank P. Hollenbeck (Purdue University) for critical reading of the manuscript. J. Barbieri (Medical College of Wisconsin) for plasmids. This work was supported by National Institutes of Health grants R56AI103168, K02AI085403 and R21AI105714 (Z.-Q.L.), 2R01GM103401 (C.D.) and National Natural Science Foundation of China grants 21305006 and 21475005 (X.L.).

**Author Contributions** J.Q. and Z.-Q.L. conceived the general ideas for this work. J.Q. and Z.-Q.L. planned, performed and interpreted experiments. Y.T. initiated the project, performed the bioinformatics analysis and determined the importance of the predicted mART motif in yeast toxicity. M.S., E.S.N., J.Q. and C.D. determined the reaction intermediates. K.Y., X.L. and E.S.N. performed mass spectrometric analyses. J.Q. and Z.-Q.L. wrote the manuscript and all authors provided editorial input.

**Author Information** Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to Z.-Q.L. (luoz@purdue.edu).

## METHODS

**Bacterial, yeast strains and plasmid construction.** *L. pneumophila* strains used in this study were derivatives of the Philadelphia 1 strain Lp02 (ref. 31) and were grown and maintained on CYE medium or in AYE broth as previously described[31]. When necessary antibiotics were included as described[31]. The Δ*sidE* strain was made by step-wise deletion of the 4 members using an established method[6]. For complementation experiments, the genes were inserted into pZL507 (ref. 32). All infections were performed with bacterial cultures grown to the post-exponential phase as judged by optical density of the cultures ($OD_{600}$ = 3.3–3.8) as well as increase of bacterial motility. For expression in mammalian cells, genes were cloned into pEGFPC1 (Clontech) or a 4×Flag vector[32]. The integrity of all constructs was verified by sequencing analysis.

**Cell culture, infection, transfection and co-immunoprecipitation.** HEK293 or 293T cells (ATCC) were cultured in Dulbecco's modified minimum Eagle's medium (DMEM) supplemented with 10% FBS. Cells grown to about 80% confluence were transfected with Lipofectamine 3000 (Life Technology) following manufacturer's instructions. U937 cells (ATCC) were differentiated into macrophages as described[33]. *D. discoideum* strains AX4 and AX4-HDEL–GFP were cultured in HL-5 medium as described earlier[34]. Strains of *L. pneumophila* used for infection were grown in AYE to post-exponential phase judged by optical density ($OD_{600}$ = 3.2–4.0) and by increase in motility. $2 \times 10^5$ *D. discoideum* cells seeded in 24-well plates were infected with an MOI of 0.05 for growth experiments and of 5 for immunostaining. In all cases, one hour after adding bacteria to cultured cells, infections were synchronized by washing the infected cells three times with warm PBS buffer. Total bacterial counts at indicated time points were determined by plating serially diluted saponin lysates onto bacterial media. To determine the development of the LCV in COS1 cells (ATCC) expressing Rab33b and its mutants, cells transfected for 14 h were infected with wild-type *L. pneumophila* and samples were fixed 14 h after bacterial uptake. Intracellular and extracellular bacteria were differentially stained with a *Legionella*-specific antibody and secondary antibodies conjugated to different fluorescence dyes. The category of LCVs was scored visually under a fluorescence microscope. All cell lines used were directly purchased from ATCC and were free of mycoplasma contamination by monthly testing using the PlasmoTest Kit (Invivogen).

For infections to determine the modification of Rab33b, HEK293 cells were transfected to express 4×Flag-Rab33b and FCγRII for 24 h with Lipofectamine 3000 (Life Technology). Bacteria of relevant *L. pneumophila* strains were opsonized with rabbit anti-*Legionella* antibodies[32] at 1:500 for 30 min before infecting the cells at an MOI of 10 for 2 h. Lysates prepared from infected cells with RIPA buffer (Thermo Fisher Scientific) were subjected to immunoprecipitation with Flag beads (Sigma-Aldrich).

To determine protein translocation by *L. pneumophila*, cells infected with the indicated bacterial strains were lysed with 0.2% saponin, which lyses membranes of mammalian cells but not of bacterial cells. The lysates were directly probed for SdeA with a specific antibody.

The secretion of SEAP was measured 24 h after cells were transfected with plasmids carrying the testing genes and pSEAP[22,35]. The alkaline phosphatase activity was determined with Tropix phosphalight System kit (Applied Biosystems) per the manufacturer's instructions.

**Yeast toxicity assays.** All yeast strains used were derived from W303 (ref. 36); yeast was grown at 30 °C in YPD medium or in appropriate amino acid dropout synthetic media with glucose or galactose at a final concentration of 2% as the sole carbon source. Yeast transformation was performed according to a standard procedure[37]. Inducible protein toxicity was assessed by the galactose-inducible promoter on pSB157 (ref. 38). SdeA or its mutant was inserted into pSB157 and the resulting plasmids were linearized before transforming into yeast strain W303 (ref. 36). Yeast strains grown in liquid selective medium containing glucose were serially diluted fivefold, and 10 μl of each dilution was spotted onto selective plates containing glucose or galactose. Plates were incubated at 30 °C for 3 days before the images were acquired.

**Protein purification.** To purify Flag-Rab33b from mammalian cells, 293T cells transfected with the indicated plasmids for 24 h were lysed with RIPA buffer. Flag-antibody-coated beads were added to cleared lysates and obtained by centrifugation at 12,000*g* for 10 min. The mixtures were incubated at 4 °C with agitation for 4 h. Unbound proteins were removed by washing the beads three times with RIPA buffer and the Flag-tagged proteins were eluted with 450 μg ml$^{-1}$ 3 × Flag peptide solution. To purify modified Rab33b from infected cells, HEK293 cells transfected to express 4×Flag-Rab33b and FCγRII were infected with wild type *L. pneumophila* for 2 h. The samples were lysed with RIPA buffer. Flag-Rab33b from the infection samples were purified followed the same protocol used for transfection samples.

Unless otherwise specified, the *E. coli* strain BL21(DE3) was used as the host for expression and purification of recombinant proteins. Rab1 was purified as GST-tagged protein, while all other proteins were purified as $His_6$-tagged proteins. pQE30–4×Flag–Rab33b was sub-cloned from the mammalian expression vector p4×Flag–Rab33b to produce $His_6$–4×Flag–Rab33b. For protein production, 30 ml of overnight culture of the *E. coli* strain harbouring the appropriate plasmid was transferred to 750 ml LB medium (ampicillin 100 μg ml$^{-1}$) and grown until $OD_{600}$ of 0.6–0.8 was reached. After adding IPTG (isopropyl thio-D-galactopyranoside) to a final concentration of 0.2 mM, the cultures were further incubated in a shaker at 18 °C for 16∼18 h. Bacterial cells were harvested by spinning at 12,000*g* and lysed by sonication in the presence of protease inhibitors. The soluble fractions were collected by centrifugation at 12,000*g* twice at 4 °C. His-tagged proteins were purified with $Ni^{2+}$-NTA beads (Qiagen), and eluted with PBS containing 300 mM imidazole; GST-Rab1 were purified with Glutathione Sepharose 4 Fast Flow beads (GE healthcare), and proteins bound to beads were eluted with 25 mM reduced glutathione in 20 mM Tris-HCl, pH 8.0, 100 mM NaCl. Purified proteins were dialysed in a buffer containing 25 mM Tris-HCl, pH 7.5, 150 mM NaCl, 5% glycerol, 1 mM DTT. To determine the potential involvement of the ions and other co-factors in the activity of SdeA, the protein was dialysed against the same buffer containing 10 mM EDTA for 14 h at 4 °C. Protein concentrations were determined by the Bradford assay. For proteins used in *in vitro* biochemical assays, extensive dialysis was performed with at least two buffer changes. The purity of proteins was larger than 95% as assessed by Coomassie brilliant blue staining.

***In vitro* ubiquitination assays.** E1, E2 and ubiquitin were obtained from Boston Biochem and were used at 100 nM for each 50-μl reaction. Ubiquitination assays were performed at 37 °C for 2 h in a reaction buffer containing 50 mM Tris-HCl (pH 7.5), 0.4 mM β-nicotinamide adenine dinucleotide (β-NAD) (Sigma-Aldrich) and 1 mM DTT. Each 50-μl reaction contains 10 μg ubiquitin, 5 μg SdeA, SdeB, SdeC, SidE or their mutant proteins and 5 μg substrates. When necessary, ATP and $Mg^{2+}$ were added to a final concentration of 2 mM and 5 mM, respectively. When needed, 50 μg of mammalian or *E. coli* lysates were added. Heat treatment of cell lysates or NAD was performed at 100 °C for 5 min. When necessary maleimide (MEM) was added to *in vitro* reactions at a final concentration of 50 μM.

**Antibodies, immunostaining and immumobloting.** Antibodies against *Legionella* and GFP were described elsewhere[32]. Antibodies specific for SdeA were prepared by injecting rabbits with purified protein (Pocono Rabbit Farm and Laboratory, Canadensis, PA) following a standard procedure used by the service provider. When necessary, antibodies were affinity-purified against the same proteins covalently coupled to an Affigel matrix (Bio-Rad) using standard protocols[39]. Cell fixation, permeabilization and immunostaining were performed as described[40]. For immunostaining, anti-*Legionella* antisera were used at 1:10,000 (ref. 32). Intracellular bacteria were distinguished from extracellular bacteria by differential immunostaining with secondary antibodies of distinct fluorescence dyes. Processed samples were inspected and scored using an Olympus IX-81 fluorescence microscope.

For immunoblotting, samples resolved by SDS–PAGE were transferred onto nitrocellulose membranes. After blocking with 5% milk, membranes were incubated with the appropriate primary antibody: anti-GFP (Sigma, cat. no. G7781), 1:10,000; anti-GST (Sigma, cat. no. G6539), 1:10,000; anti-Flag (Sigma, F1804), 1:2,000; anti-ICDH, 1:10,000; anti-PGK (Life Technology, cat. no. 459250), 1:3,000; anti-SdeA, 1:10,000; anti-SidC[6], 1:10,000; anti-Ub (Santa cruz, cat. no. sc-8017), 1:1,000; anti-His (Sigma, cat. no. H1029), 1:10,000. Tubulin (DSHB, E7), 1:10,000. Membranes were incubated with an appropriate IRDye infrared secondary antibody (Li-Cor's Biosciences Lincoln, Nebraska, USA) and the signals were obtained by using the Odyssey infrared imaging system.

**GTP loading assay.** For $^{35}$SγGTP incorporation assays, 20 μg of 4×Flag-Rab33b was loaded with unlabelled GDP (5 mM) before ubiquitination as described[22]. GDP loaded 4×Flag-Rab33b was used for ubiquitination assays in the presence of either SdeA (10 μg) or $SdeA_{E/A}$ (10 μg) for 2 h at 37 °C. 20% of the samples were withdrawn to test for the extent of ubiquitination of 4×Flag-Rab33b by SDS–PAGE and Coomassie staining. Ubiquitinated or non-ubiquitinated 4×Flag-Rab33b was incubated in 50 μl nucleotide exchange buffer containing 25 mM Tris-HCl (pH 7.5), 50 mM NaCl, 5 mM $MgCl_2$, and 0.1 mM EDTA with 5 μCi $^{35}$SγGTP (Perkin-Elmer). GTP-loading reactions were performed at 22 °C. Aliquots of reactions were withdrawn at indicated time points, passed through nitrocellulose membrane filters (Hawp02500; Millipore) and placed onto a vacuum platform attached to a waste liquid container. Membranes were washed three times using the exchange buffer to remove the free nucleotides, and were then transferred into scintillation vials containing 8 ml scintillation fluid (Beckman). Incorporated $^{35}$SγGTP was detected by a scintillation counter at 1 min per count.

**GTPase assay.** 20 μg of 4×Flag-Rab33b was used for ubiquitination assays in the presence of either SdeA (10 μg) or $SdeA_{E/A}$ (10 μg) for 2 h before 5 μCi of $^{32}$PγGTP (Perkin-Elmer) was added to the reactions. Nucleotide loading was performed at 22 °C for 30 min. Aliquots of the reactions were withdrawn and passed through

membranes as described in the GTP loading assay. The reading of these aliquots served as starting points for different reactions. Samples withdrawn at later time points were measured for $^{32}$P$\gamma$GTP and retained by 4×Flag-Rab33b-bound with a scintillation counter. The GTP hydrolysis index was calculated by dividing the readings obtained in later time points by the values of the starting point.

**ADP-ribosylation assay.** 5 μg of SdeA or SdeA$_{E/A}$ was incubated with 5 μg of GST-Rab1, 4×Flag-Rab33b or 100 μg of 293T cell lysate in the presence of 10 mM Tris-HCl (pH 7.5), 20 mM NaCl. 5 μCi of $^{32}$P-α-NAD (Perkin-Elmer) was added to each reaction. ADP-ribosylation assays were performed at 22 °C for 1 h and were stopped by adding 5 × SDS loading buffer. A reaction containing ExoS$_{78-453}$ (200 ng), FAS (factor activating ExoS) (2 μg), Rab5 (5 μg) or 293T cell lysates (100 μg) was used as positive control. The incorporation of $^{32}$P-α-ADPR into proteins was detected by autoradiography.

**Detection of reaction intermediates by $^{32}$P-labelled ATP and NAD.** To detect the ubiquitin intermediate, 5 μg of SdeA or SdeA$_{519-1100}$ was incubated with 10-μg GST-ubiquitin, GST-ubiquitin$_{R42A}$ or GST in the presence of $^{32}$P-α-NAD (5 μCi) in a reaction buffer containing 50 mM Tris-HCl (pH 7.5). The reaction was performed at 37 °C for 6 h and stopped by adding 5 × SDS loading buffer. A reaction containing the E1 activating enzyme (1 μg), GST-ubiquitin or GST (10 μg), $^{32}$P-α-ATP (5 μCi) in the presence of 50 mM Tris-HCl (pH 7.5) and 2 mM $MgCl_2$ was used as a positive control. The $^{32}$P-labelled intermediates were detected by autoradiography.

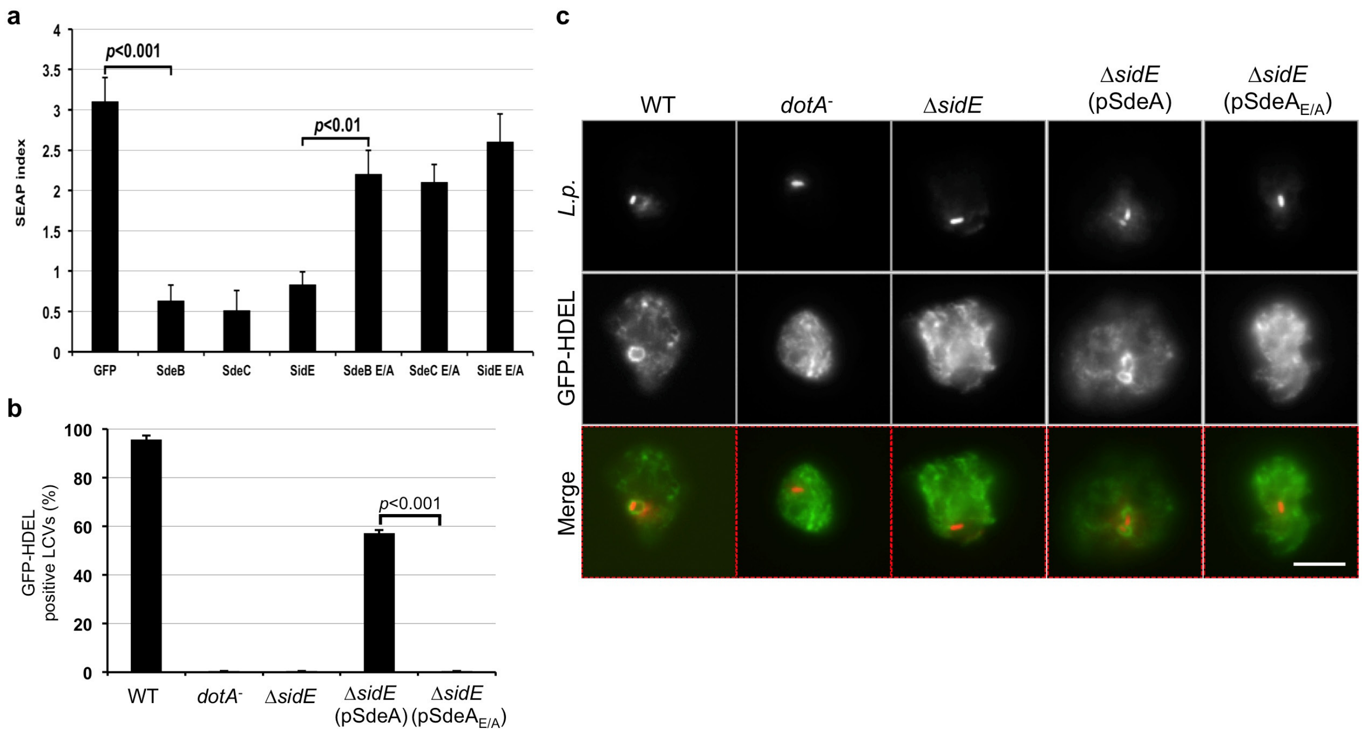
**Detection of reaction intermediates.** To detect AMP generated in reactions catalysed by SdeA, reactions were set up with 50 μg SdeA$_{178-1000}$, 10 mM NAD and 450 μg ubiquitin in reaction buffer (50 mM Tris pH 7.6, 50 mM NaCl, 1 mM DTT) and allowed to react for 2 h at 22 °C. To detect all reaction intermediates, a reaction was set up with 100 μg SdeA$_{178-1000}$, 1 mM NAD and 100 μg ubiquitin in reaction buffer (50 mM Tris pH 7.6, 50 mM NaCl, 1 mM DTT) and allowed to react for 16 h at 22 °C. The reaction was then separated on an Agilent C8 column using a Waters 600 HPLC system with a linear gradient of 0–5% (v/v) acetonitrile in water over 25 min at 1 ml per minute. The intermediates were detected with a Waters 2487 dual wavelength detection system with wavelengths set to 260 nm and 280 nm. The mixture was then directly analysed with a Waters micromass ZQ spectrometer in negative electrospray ionization mode. The detection range was set from 100–700 ($m/z$) with a scans at 1 s intervals. Standard samples of AMP, ADP, NMN, and nicotinamide were set up in parallel and analysed following the same method to determine the elution profile of each possible intermediate.

For experiments using SdeA$_{519-1100}$ defective in autoubiquitination, 50 μg SdeA$_{519-1100}$ was incubated with 15 μg ubiquitin and 1 mM NAD in reaction buffer (50 mM Tris pH 7.6, 50 mM NaCl, 1 mM DTT) at 22 °C for 18 h. The reaction was then applied directly to an Agilent C8 column on a Waters 600 HPLC system. The products of the reaction were separated with a linear gradient of 0–5% (v/v) acetonitrile in water with a flow rate of 1 ml per min over 25 min. The products were detected with a Waters 2487 dual wavelength detection system set to 260 nm and 280 nm. Controls used were 1 mM solutions containing only NAD, nicotinamide or AMP.

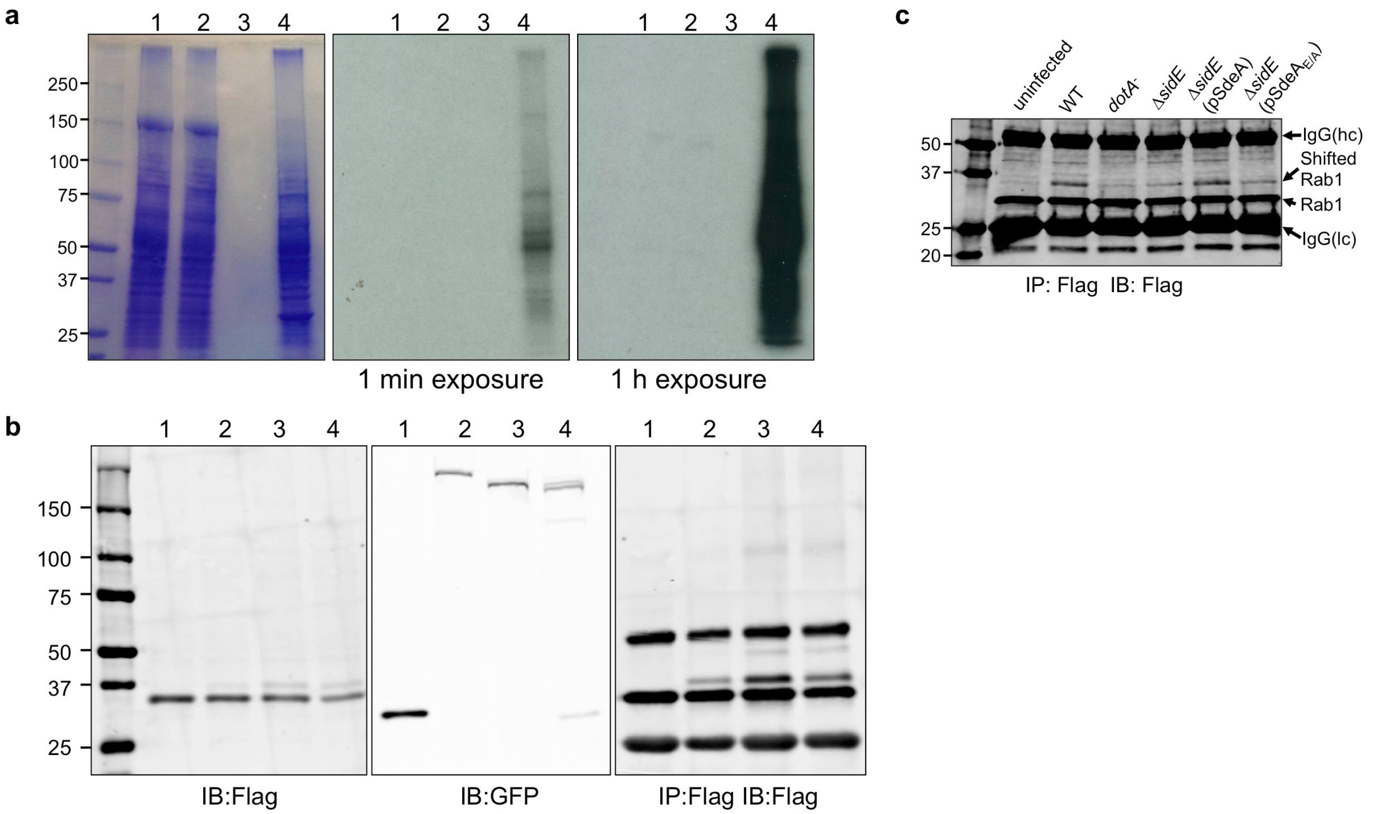
Samples for mass spectrometric analysis were obtained by using His$_6$–ubiquitin in reactions containing SdeA$_{519-1100}$ and NAD for 2 h, SdeA$_{519-1100}$ and other components were removed by $Ni^{2+}$ beads chromatography. Eluted proteins were separated in SDS–PAGE and the band corresponding ubiquitin was excised and digested with trypsin. Resulting peptides were analysed in a NanoAcquity nano-HPLC system (Waters) by loading peptides into a trap column (5 cm × 150 μm i.d. column packed in-lab with 5 μm Jupiter C18 stationary phase) and separated in a 40 cm × 75 μm i.d. column packed in-lab with 3 μm Jupiter C18 stationary phase. The elution was carried out at 300 nl per min with the following gradient: 0–8% B solvent in 2 min, 8–20% B in 18 min, 12–30% B 55 min, 30–45% B in 22 and 97–100% B in 3 min, before holding for 10 min at 100% B. Eluting peptides were introduced to the mass spectrometer (Q-Exactive HF, Thermo Fisher Scientific) using electrospray ionization and mass spectra were collected from 400–2,000 $m/z$ with 100,000 resolution at $m/z$ 400. HCD tandem-mass spectra were collected by data-dependent acquisition of the 12 most intense ions using normalized collision energy of 30%. A dynamic exclusion time of 45 s was used to discriminate against previously analysed ions. Spectra were analysed manually by *de novo* sequencing.

**Data quantitation and statistical analyses.** Student's $t$-test (two-sided) was used to compare the mean levels between two groups each with at least three independent samples. No statistical methods were used to predetermine sample size.

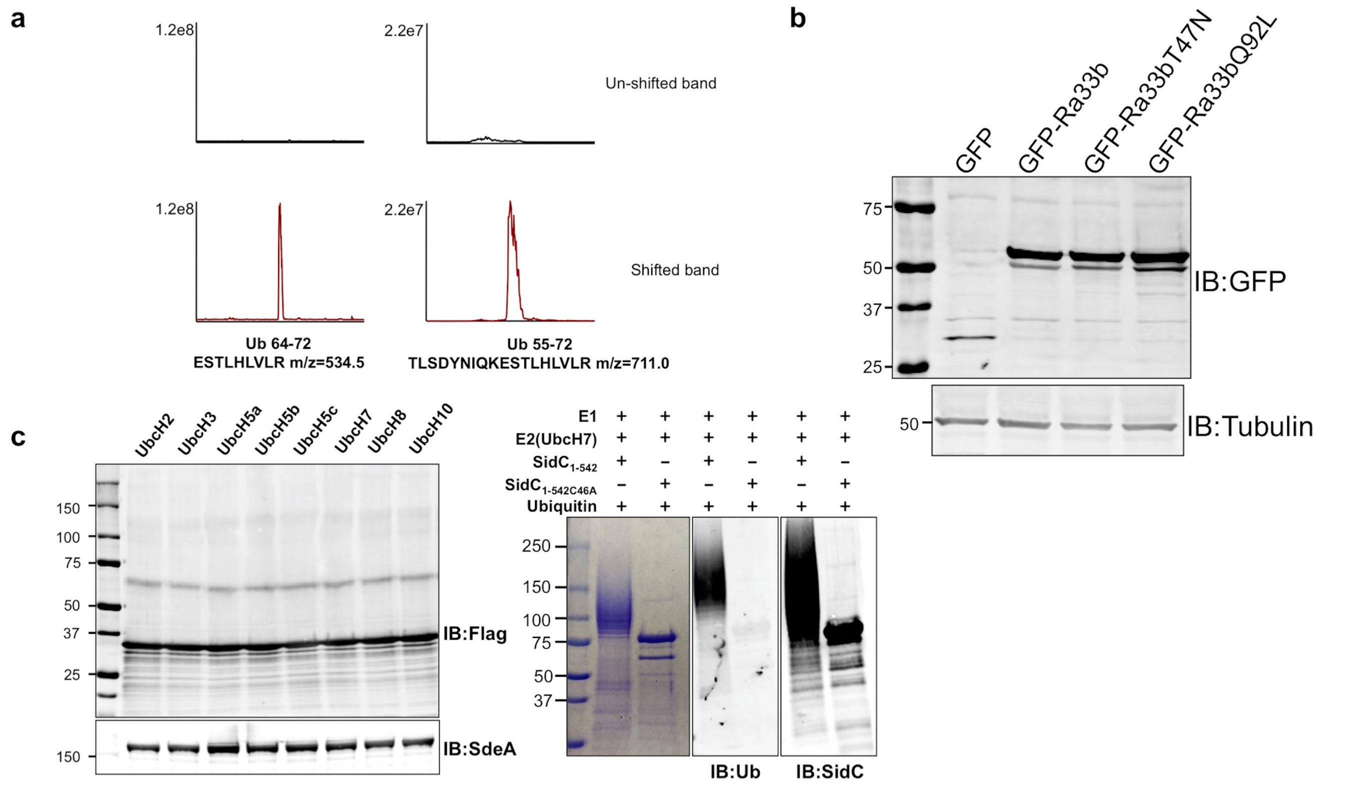

31. Berger, K. H. & Isberg, R. R. Two distinct defects in intracellular growth complemented by a single genetic locus in *Legionella pneumophila*. *Mol. Microbiol.* **7,** 7–19 (1993).
32. Xu, L. *et al.* Inhibition of host vacuolar H+-ATPase activity by a *Legionella pneumophila* effector. *PLoS Pathog.* **6,** e1000822 (2010).
33. Tilney, L. G., Harb, O. S., Connelly, P. S., Robinson, C. G. & Roy, C. R. How the parasitic bacterium *Legionella pneumophila* modifies its phagosome and transforms it into rough ER: implications for conversion of plasma membrane to the ER membrane. *J. Cell Sci.* **114,** 4637–4650 (2001).
34. Li, Z., Solomon, J. M. & Isberg, R. R. *Dictyostelium discoideum* strains lacking the RtoA protein are defective for maturation of the *Legionella pneumophila* replication vacuole. *Cell. Microbiol.* **7,** 431–442 (2005).
35. Pan, X., Luhrmann, A., Satoh, A., Laskowski-Arce, M. A. & Roy, C. R. Ankyrin repeat proteins comprise a diverse family of bacterial type IV effectors. *Science* **320,** 1651–1654 (2008).
36. Fan, H. Y., Cheng, K. K. & Klein, H. L. Mutations in the RNA polymerase II transcription machinery suppress the hyperrecombination mutant Hpr1Δ of *Saccharomyces cerevisiae*. *Genetics* **142,** 749–759 (1996).
37. Gietz, R. D., Schiestl, R. H., Willems, A. R. & Woods, R. A. Studies on the transformation of intact yeast cells by the LiAc/SS-DNA/PEG procedure. *Yeast* **11,** 355–360 (1995).
38. Fazzio, T. G. & Tsukiyama, T. Chromatin remodeling *in vivo:* evidence for a nucleosome sliding mechanism. *Mol. Cell* **12,** 1333–1340 (2003).
39. Duménil, G. & Isberg, R. R. The *Legionella pneumophila* IcmR protein exhibits chaperone activity for IcmQ by preventing its participation in high-molecular-weight complexes. *Mol. Microbiol.* **40,** 1113–1127 (2001).
40. Conover, G. M., Derre, I., Vogel, J. P. & Isberg, R. R. The *Legionella pneumophila* LidA protein: a translocated substrate of the Dot/Icm system associated with maintenance of bacterial integrity. *Mol. Microbiol.* **48,** 305–321 (2003).



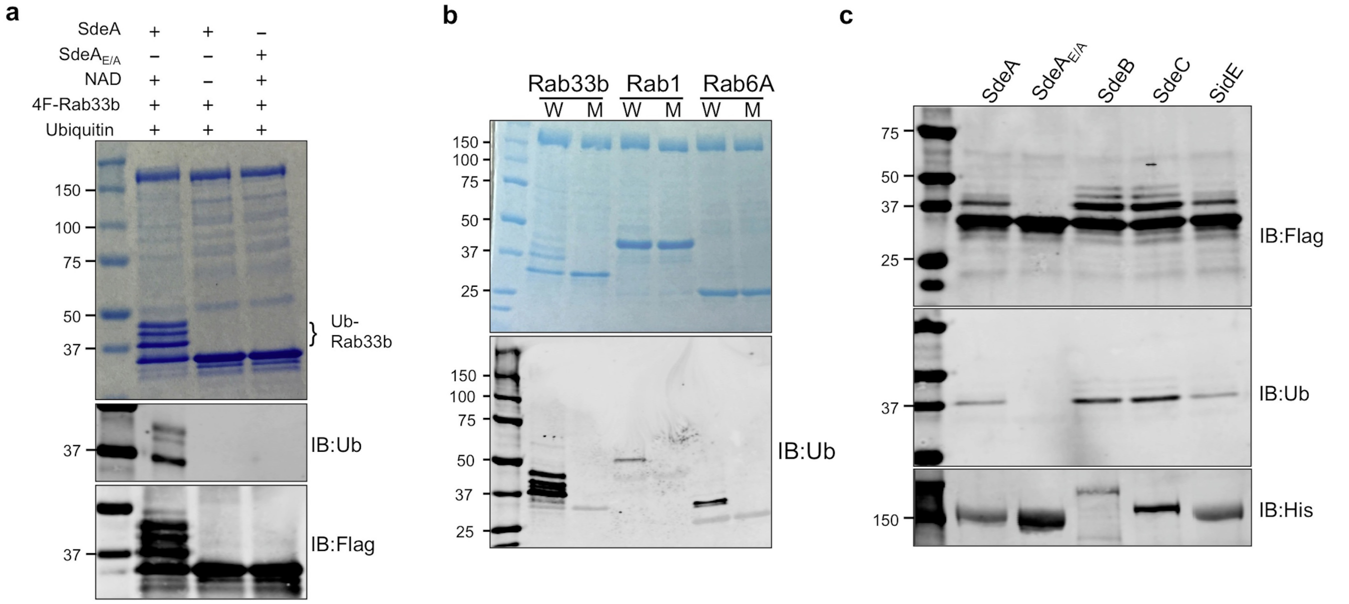

**Extended Data Figure 1 | Inhibition of the secretion of SEAP by SidE, SdeB and SdeC and the recruitment of an ER marker by the *L. pneumophila* mutant lacking the SidE family. a**, GFP fusions of the indicated proteins were co-expressed with SEAP in 293T cells for 24 h. The SEAP index was determined by measuring alkaline phosphatase activity in culture supernatant or in cells. Similar results were obtained in three independent experiments, and data shown are from one representative experiment done in triplicate. Note that mutations in the putative mART motif abolished the inhibitory effects. Error bars represent s.e.m. ($n = 3$). **b**, Quantitation of the vacuoles positive for GFP–HDEL. The indicated bacterial strains were used to infect a line of *D. discoideum* stably expressing GFP fusion to the ER retention signal HDEL and the recruitment of the GFP–HDEL signal to the phagosome was evaluated 10 h after infection. At least 150 phagosomes were scored in each sample done in triplicate. Results shown are from one representative experiment done in triplicate and similar results were obtained from three independent experiments. Error bars represent s.e.m. ($n = 3$). **c**, Representative images of *L. pneumophila* phagosomes associated with GFP–HDEL. Images are from one representative of three independent experiments with similar results. Scale bar, 5 μm.



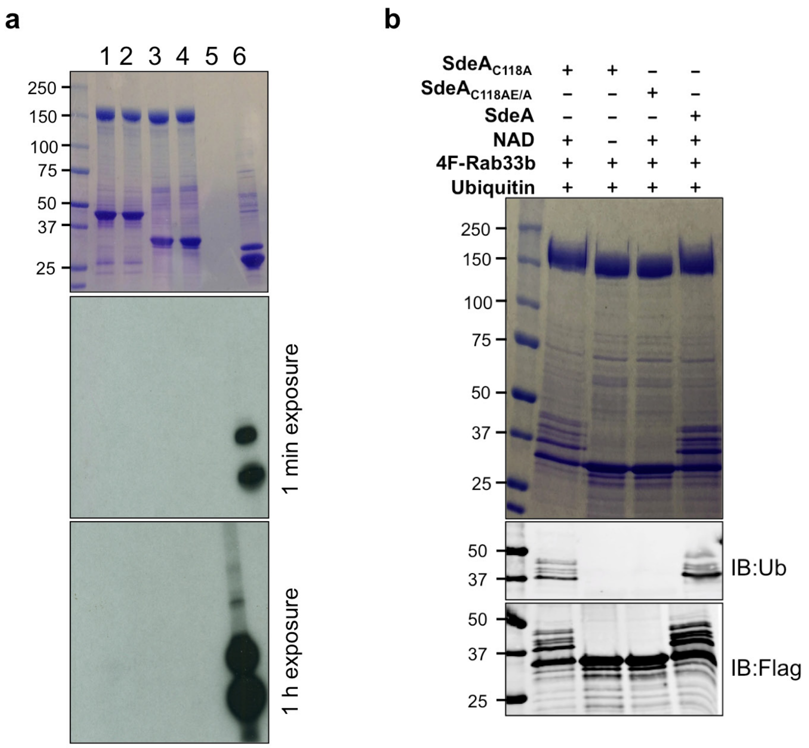

**Extended Data Figure 2 | SdeA does not ADP-ribosylate mammalian proteins, the modification of Rab33b by other members of the SidE family and SdeA-mediated post-translational modification of Rab1 during bacterial infection. a**, SdeA, $\text{SdeA}_{E/A}$ or ExoS and 5 μCi $^{32}$P-NAD were added to 100 μg total protein of 293T cells. After incubation at 22 °C for 1 h, samples were separated by SDS–PAGE. Gels were stained with Coomassie brilliant blue (left panel) and then by autoradiography for the indicated time duration (middle and right panels). In samples receiving SdeA, no ADP-ribosylation signal was detected in many experiments performed in various reaction conditions. Lane 1: $^{32}$P-α-NAD + SdeA + 293T lysates; lane 2: $^{32}$P-α-NAD + $\text{SdeA}_{E/A}$ + 293T lysates; lane 3: no sample; lane 4: $^{32}$P-α-NAD + $\text{ExoS}_{78\text{-}453}$ + FAS + 293T lysates. **b**, Flag-tagged Rab33b was co-expressed with GFP-tagged testing proteins in 293T cells for 24 h. Cell lysates were subjected to immunoprecipitation with Flag beads and the precipitated products were probed with the Flag antibody (right panel). 5% of each lysate was probed for the expression of Rab33b (left panel) or for GFP fusions (middle panel). Proteins used: 1, GFP; 2, GFP–$\text{SdeB}_{1\text{-}1751}$; 3, GFP–SdeC; 4, GFP–SidE. **c**, 293T cells transfected to express Flag–Rab1 were infected with the indicated *L. pneumophila* strains for 2 h and the Rab1 enriched by immunoprecipitation was probed by immunoblotting. For all panels, similar results were obtained from three experiments. **a**–**c**, Uncropped blots and autoradiograph images are shown in Supplementary Fig. 1.



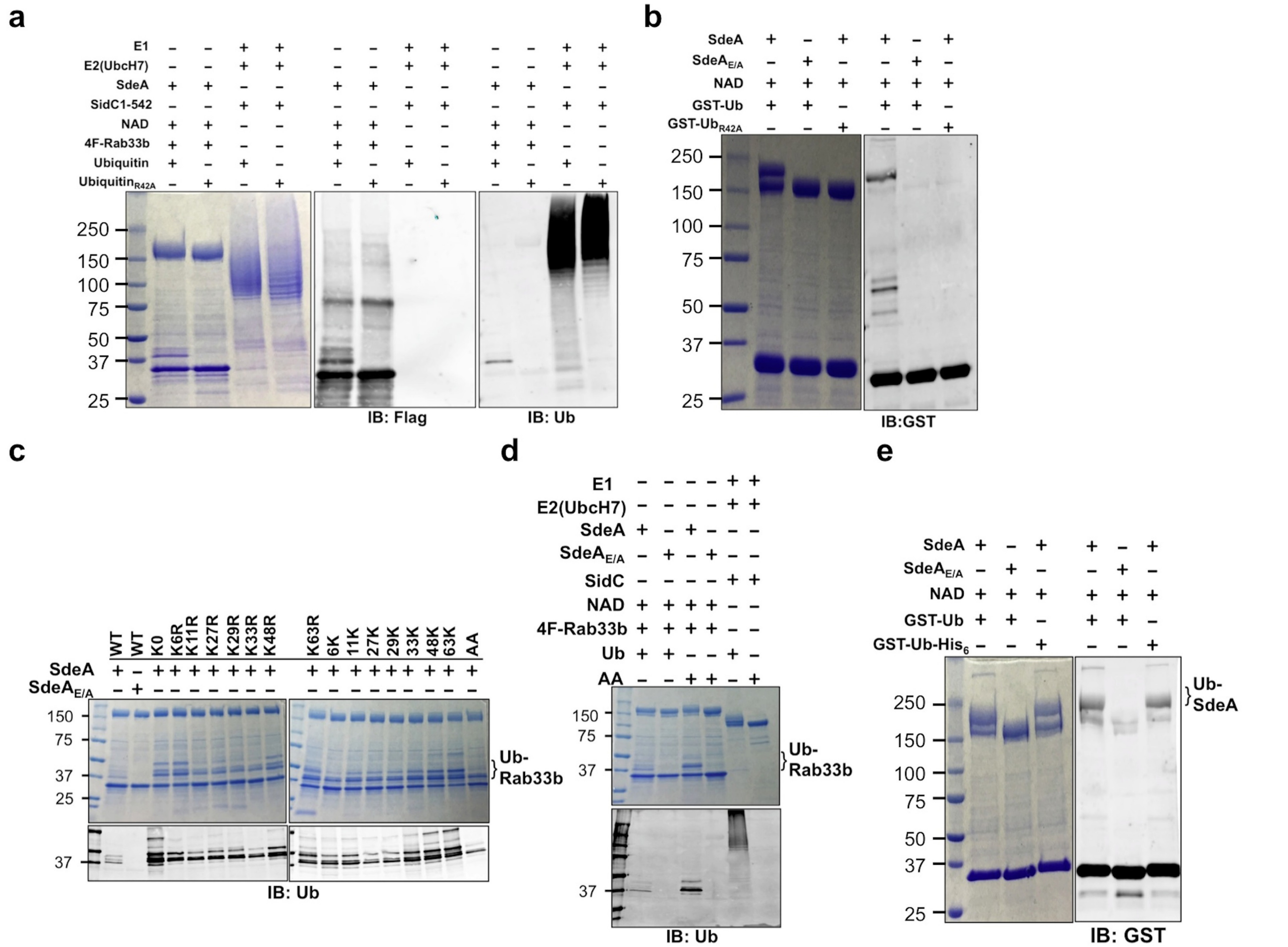

**Extended Data Figure 3 | The extracted ion chromatograms of ubiquitin tryptic fragments detected by mass spectrometry, expression of Rab33b and its mutants in COS1 cells, and *in vitro* ubiquitination of Rab33b by SdeA with E1 and a series of E2 proteins. a**, Proteins in bands corresponding to normal (upper panel) or shifted (lower panel) Rab33b were digested with trypsin and the resulting protein fragments were identified by mass spectrometry. Note that the ubiquitin tryptic fragments are present only in the shifted band of higher molecular mass. **b**, COS1 cells were transfected with GFP or GFP fusion of Rab33b or its mutants for 14 h. Total cell lysates resolved by SDS–PAGE were probed with a GFP-specific antibody. Tubulin was detected as a loading control. **c**, Reactions containing E1 and the indicated E2 proteins were allowed to proceed at 37 °C for 2 h. Proteins in the reactions were resolved by SDS–PAGE followed by immunoblotting to detect ubiquitinated proteins with higher molecular mass (left panel). SdeA in the reaction was detected with specific antibodies by using 10% of the reactions (lower panel). Control reactions with wild-type *Legionella* E3 ligase SidC$_{1-542}$ and its enzymatically inactive mutant SidC$_{1-542C46A}$ with E1 and the E2 UbcH7 were established to monitor the activity of E1 (right panel). Note the robust self-ubiquitination of SidC$_{1-542}$ (second lane right panel). Results in **a** are representative of three experiments with similar results; **b** and **c** are a representative of two and five independent experiments, respectively. **b**, **c**, Uncropped blots are shown in Supplementary Fig. 1.



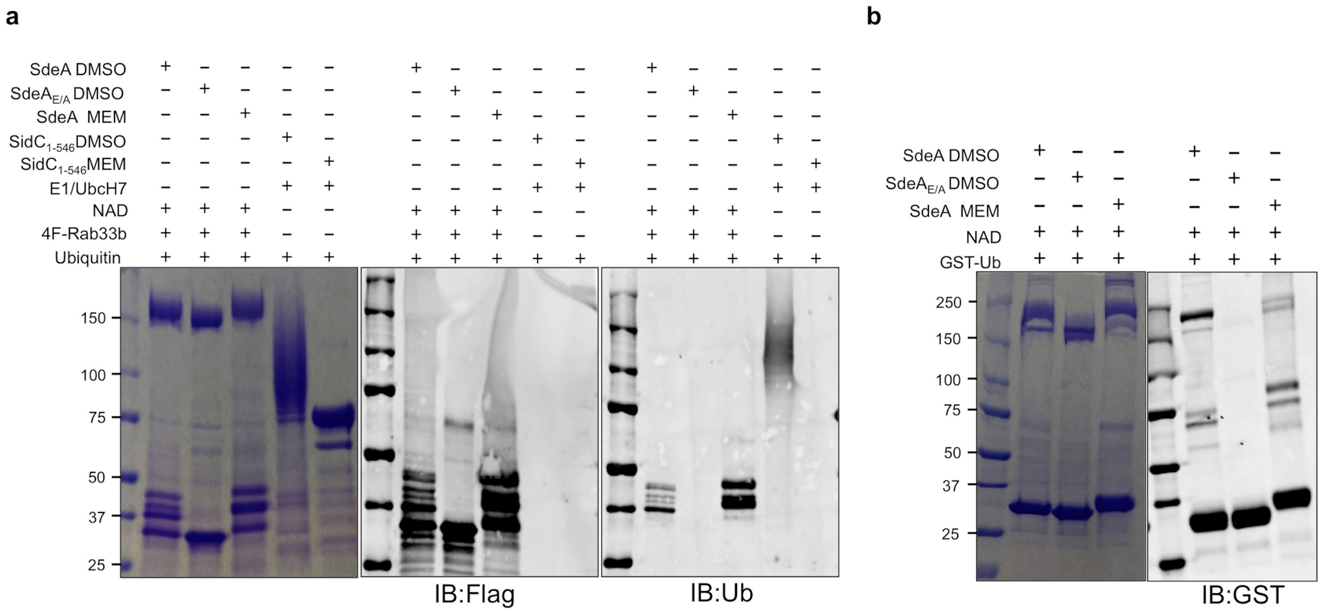

**Extended Data Figure 4 | The activity of EDTA-dialysed SdeA and other members of the SidE family.** **a**, SdeA or $SdeA_{E/A}$ dialysed against a buffer containing 10 mM EDTA was used for *in vitro* ubiquitination of Rab33b. Reactions were allowed to proceed for 2 h at 37 °C. Samples resolved by SDS–PAGE were detected by Coomassie staining (upper panel), by immunoblotting with antibodies specific for ubiquitin (middle panel) or for the Flag tag (lower panel). Note that the addition of exogenous NAD is sufficient to allow SdeA-mediated ubiquitination of Rab33b (lane 2). **b**, *In vitro* ubiquitination of Rabs by SdeA. Reactions containing indicated proteins and NAD were allowed to proceed for 2 h at 37 °C. After SDS–PAGE, ubiquitinated proteins were detected by staining 50% of the reactions resolved by SDS–PAGE with Coomassie (upper panel) or by immunoblotting with antibodies specific for ubiquitin (lower panel). Similar results were obtained from two experiments. **c**, *In vitro* ubiquitination of Rab33b by SidE, $SdeB_{1-1751}$ and SdeC. Indicated testing proteins were incubated with NAD, ubiquitin and Flag–Rab33b for 2 h at 37 °C. Proteins resolved by SDS–PAGE were detected by antibodies specific for Flag (upper panel) or for ubiquitin (middle panel). $His_6$-tagged SdeA, $SdeB_{1-1751}$ and SdeC and $SdeA_{E/A}$ used in the reactions were probed 10% of the proteins with an antibody against His (lower panel). Similar results were obtained from two independent experiments. **a**–**c**, Uncropped blots are shown in Supplementary Fig. 1.



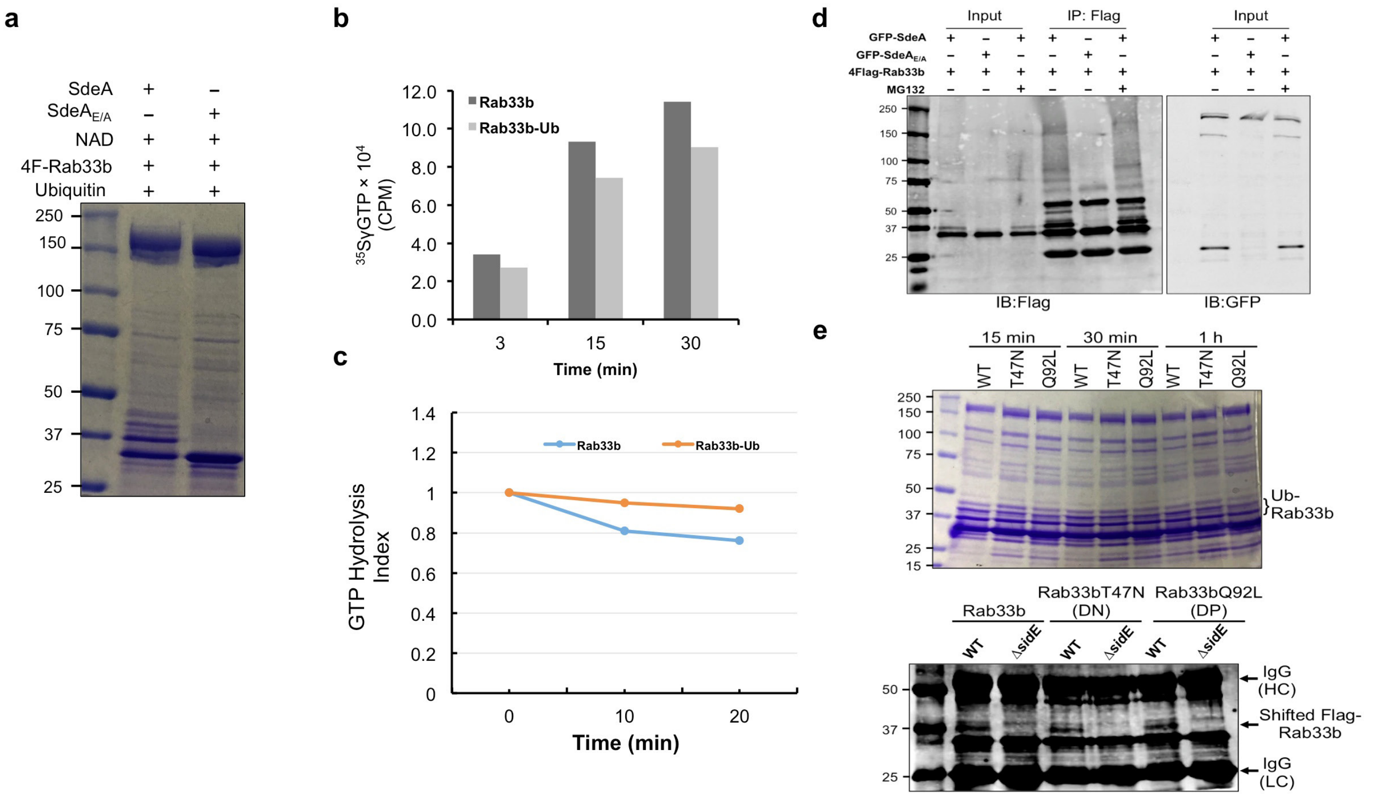

**Extended Data Figure 5 | SdeA does not detectably ADP-ribosylate Rab33b or Rab1 and the deubiquitinase (DUB) activity of SdeA does not interfere with its ubiquitin-conjugation activity. a**, 5 μg of SdeA or $SdeA_{E/A}$ were incubated with 5 μg of GST–Rab1, 4×Flag–Rab33b and 5 μCi of $^{32}$P-α-NAD. A reaction containing 200 ng of $ExoS_{78-453}$, 2 μg of FAS and 5 μg Rab5 was established as a positive control. All reactions were allowed to proceed for 1 h at 22 °C before being terminated by adding 5 × SDS loading buffer. Samples resolved by SDS–PAGE were detected by Coomassie staining (upper panel) and then by autoradiography (middle and lower panels). Lane 1: $^{32}$P-α-NAD + SdeA + GST–Rab1; lane 2: $^{32}$P-α-NAD + $SdeA_{E/A}$ + GST–Rab1; lane 3: $^{32}$P-α-NAD + SdeA + 4×Flag–Rab33b; lane 4: $^{32}$P-α-NAD + $SdeA_{E/A}$ + 4×Flag–Rab33b; lane 5: no sample; lane 6: $^{32}$P-α-NAD + $ExoS_{78-453}$ + FAS + Rab5. Note the strong ADP-ribosylation signals in the reaction with $ExoS_{78-453}$ (lane 6). **b**, SdeA, its mutants $SdeA_{C118A}$ or $SdeA_{C118AE/A}$ was used for *in vitro* NAD-dependent ubiquitination of Rab33b. Reactions containing the indicated components were allowed to proceed for 2 h at 37 °C before being terminated with SDS sample buffer. Samples resolved by SDS–PAGE were probed by Coomassie staining (upper panel) or by immunoblotting with antibody specific for ubiquitin (middle panel) or for the Flag tag (lower panel). **c**, Reactions containing GST–ubiquitin were similarly established to detect self-ubiquitination by SdeA. Note that SdeA and $SdeA_{C118A}$ exhibited similar activity in these reactions. Data in all panels are one representative of two independent experiments with similar results. **a**–**c**, Uncropped blots and autoradiograph images are shown in Supplementary Fig. 1.



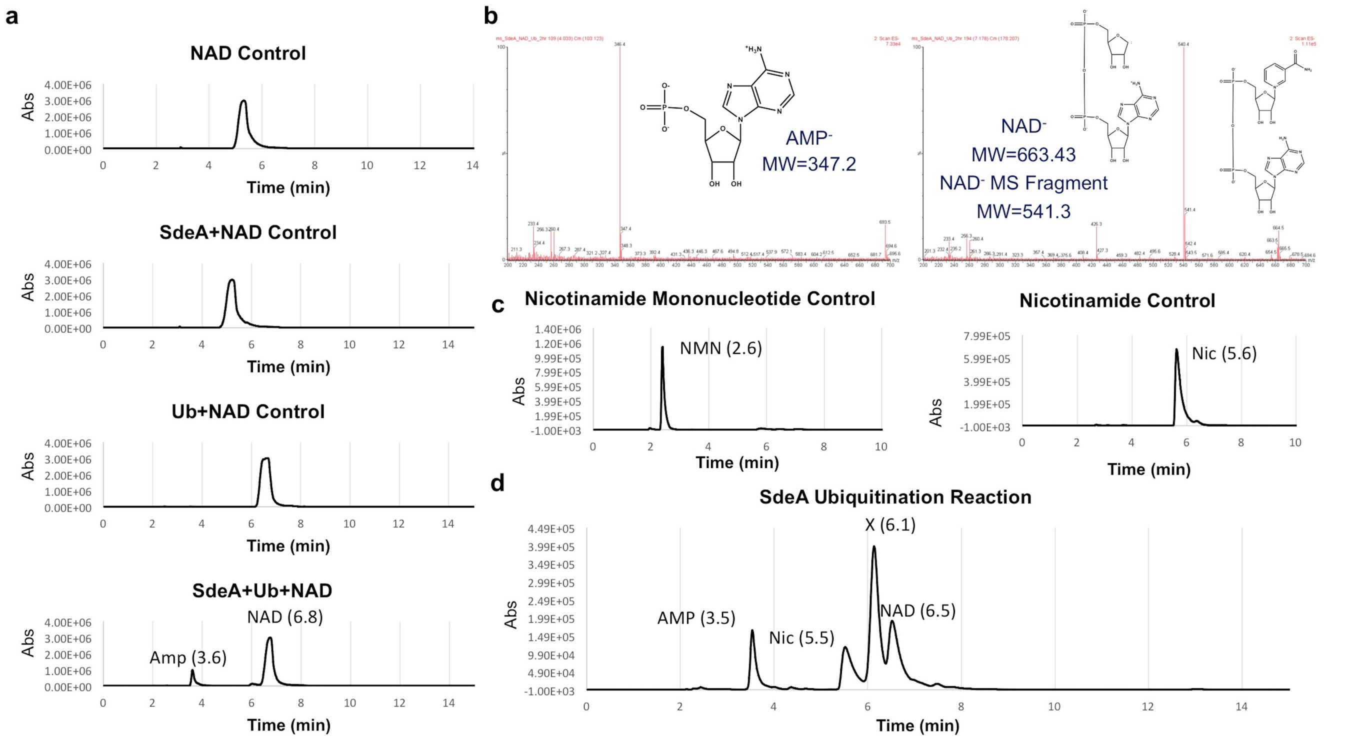

**Extended Data Figure 6 | The reactivity of ubiquitin mutants in SdeA-mediated ubiquitination. a**, Arg42 in ubiquitin is important for SdeA-mediated ubiquitination. Ubiquitin or ubiquitin$_{R42A}$ was included in reactions catalysed by SdeA or the bacterial E3 ubiquitin ligase SidC (E1 and the E2 UbcH7 were added in the latter category of reactions). After allowing the reaction to proceed for 2 h at 37 °C. Samples separated by SDS–PAGE were probed with antibody against the Flag tag (on Rab33b) (middle panel) or ubiquitin (right panel). Note that ubiquitin$_{R42A}$ can be used by ubiquitination catalysed by SidC but not SdeA. **b**, GST–ubiquitin$_{R42A}$ cannot be used for self-ubiquitination by SdeA. GST–ubiquitin or GST–ubiquitin$_{R42A}$ was used in reactions with SdeA or SdeA$_{E/A.}$ Self-modification was detected by the shift of SdeA detected by Coomassie staining (left panel) or by immunoblotting with a GST-specific antibody (right panel). **c**, The lysine residues or the carboxyl terminus of ubiquitin is not important for SdeA-catalysed Rab33b ubiquitination. Reactions containing SdeA or SdeA$_{E/A}$, NAD, Flag-Rab33b and the indicated ubiquitin mutants were allowed to proceed for 2 h at 37 °C. Proteins were detected by Coomassie staining (upper panel) or probed by immunoblotting with antibody against ubiquitin. **d**, Utilization of the ubiquitin di-glycine mutant by different ligases. Reactions with indicated components were allowed to proceed for 2 h at 37 °C. Proteins resolved by SDS–PAGE were detected by staining (upper panel) or by immunoblotting with antibodies specific to ubiquitin (lower panel). Note that the wild type but not the di-glycine ubiquitin mutant (AA) can be conjugated to proteins in a reaction containing E1 and E2 and the bacterial E3 ligase SidC (Lanes 6 and 7). This di-glycine mutant (AA) can still be attached to Rab33b by SdeA (Lane 4). **e**, Addition of 6 histidine residues to the carboxyl end of ubiquitin did not affect SdeA-mediated ubiquitination. Reactions containing the indicated components were established and allowed to proceed for 2 h at 37 °C. SDS–PAGE resolved samples were probed by Coomassie staining (left panel) or by immunoblotting with a GST-specific antibody (right panel). The data in all panels are one representative of three independent experiments with similar results. **a**–**e**, Uncropped blots are shown in Supplementary Fig. 1.



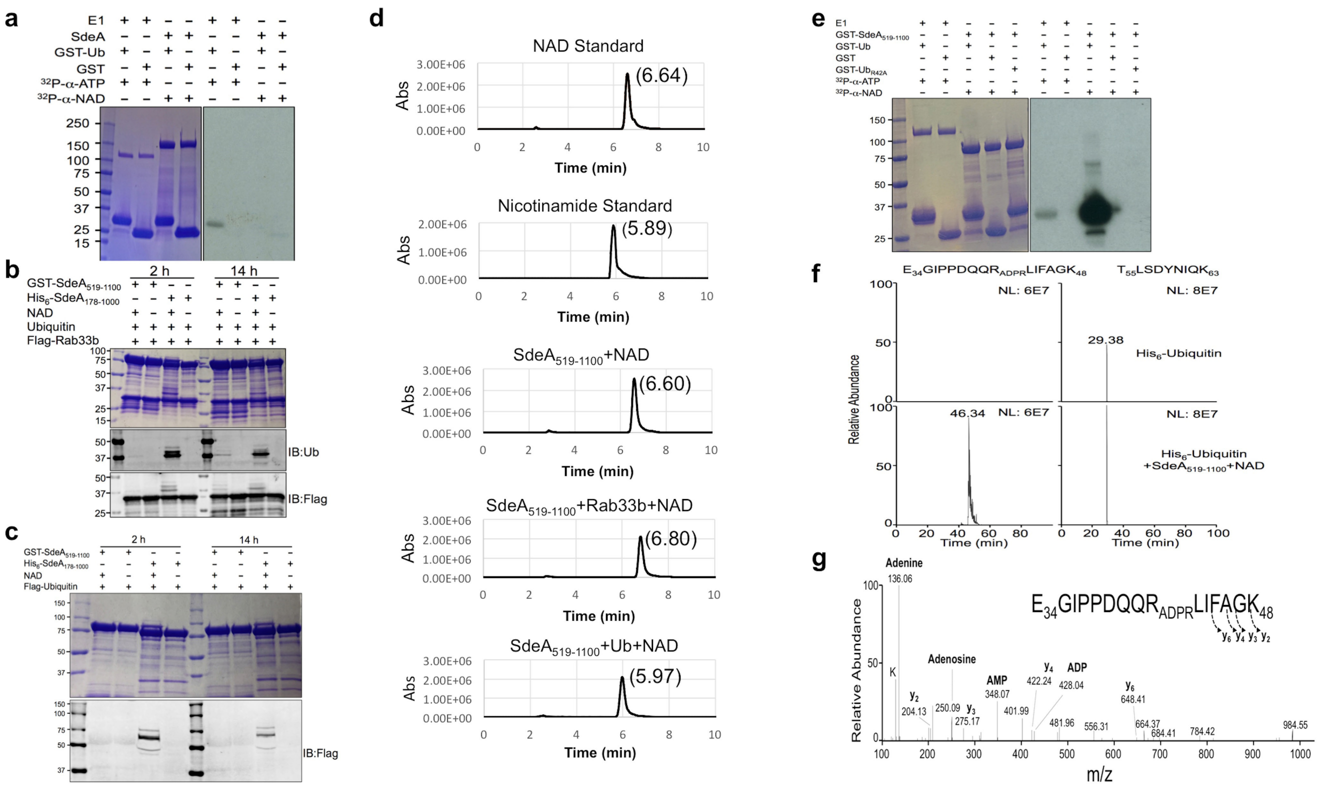

**Extended Data Figure 7 | Ubiquitination catalysed by SdeA is insensitive to the cysteine modifying agent maleimide. a**, Ubiquitination reactions by SdeA or SidC together with E1 and E2 were established; maleimide was added to 50 μM to a subset of these reactions. After incubation at 37 °C for 2 h, ubiquitination was detected by Coomassie staining (left panel) or by immunoblotting with the Flag- (middle panel) or ubiquitin-specific (right) antibody. Note that maleimide completely inhibits ubiquitination in the reaction catalysed by SidC, E1 and its cognate and E2 (lane 6) but does not affect the activity of SdeA (lane 4). **b**, Maleimide does not affect self-ubiquitination of SdeA. Reactions containing the indicated components were established and the modification of SdeA was probed by Coomassie staining (left panel) or by immunoblotting with the GST-specific antibody (right panel). For all panels, similar results were obtained from four independent experiments. **a**, **b**, Uncropped blots are shown in Supplementary Fig. 1.




**Extended Data Figure 8 | SdeA-mediated ubiquitination affects the activity but not stability of Rab33b and SdeA ubiquitinates Rab33b independently of its nucleotide binding status. a**, Evaluation of the ubiquitinated Rab33b. 4×Flag–Rab33b was loaded with unlabelled GDP (5 mM) before ubiquitination reaction. GDP-loaded Rab33b was subjected to ubiquitination by SdeA or $SdeA_{E/A}$ for 2 h at 37 °C; 20% of the samples were withdrawn to determine the extent of ubiquitination by Coomassie staining. **b**, Ubiquitination affected the GTP loading activity of Rab33b. Ubiquitinated or non-ubiquitinated 4×Flag–Rab33b was incubated in 50 μl nucleotide exchange buffer containing 5 μCi $^{35}S\gamma$GTP at 22 °C. Aliquots of reactions were withdrawn at indicated time points and passed through nitrocellulose membrane filters. Membranes were washed for three times using exchange buffer before being transferred into scintillation vials containing scintillation fluid to detect incorporated $^{35}S\gamma$GTP with a scintillation counter. **c**, Ubiquitination affected the GTPase activity of Rab33b. Samples withdrawn from Ub~Rab33b or Rab33b loaded with $^{32}P\gamma$GTP were measured for the associated radioactivity to set as the starting point. Equal volumes of samples were withdrawn at the indicated time points to monitor intrinsic GTP hydrolysis. The GTP hydrolysis index was calculated by dividing the readings obtained in later time points by the values of the starting point. Similar results (**a**–**c**) were obtained in three independent experiments and the data shown were from one representative experiment. **d**, SdeA-mediated ubiquitination does not lead to degradation of Rab33b. GFP fusion of SdeA or $SdeA_{E/A}$ was co-transfected with Rab33b for 14 h. The proteasome inhibitor MG132 (10 μM) was added to one of the SdeA samples. The levels of Rab33b were detected by immunoblotting following immunoprecipitation with the Flag-specific antibody. Note that the addition of MG132 does not affect the level of modified Rab33b in samples co-transfected with SdeA. Similar results were obtained from two independent experiments. **e**, The nucleotide binding status of Rab33b does not affect its suitability as substrate in SdeA-mediated ubiquitination. Equal amounts of Rab33b, its dominant negative mutant Rab33b(T47N), or the dominant positive mutant Rab33b(Q92L) was incubated with SdeA. Samples withdrawn at the indicated time points were detected for ubiquitination by Coomassie staining (upper panel); 293T cells transfected to express these mutants were infected the indicated *L. pneumophila* strains and ubiquitinated Rab33b or its mutants were probed by molecular mass shift in Rab33b obtained by immunoprecipitation (lower panel). Data in this panel are one representative of two independent experiments with similar results. **a**, **d**, **e**, Uncropped blots and gel images are shown in Supplementary Fig. 1.




**Extended Data Figure 9 | Detection of the reaction intermediates in SdeA-catalysed ubiquitination. a**, Controls were analysed by HPLC of NAD alone and in the presence of SdeA, Ub, and SdeA and Ub. In these reactions, AMP and NAD were identified with retention times of 3.6 and 6.8 min, respectively. **b**, Both AMP (left) and NAD (right) were additionally identified by ESI mass spectrometry. Both NAD and a product in which the nicotinamide group has been lost were observed in these experiments. **c**, To determine whether other fragments are generated in this reaction, retention time for nicotinamide mononucleotide (NMN, left) and nicotinamide (Nic, right) was determined by HPLC to be 5.6 and 2.6 min respectively. **d**, To identify additional components, a reaction was set up and the individual components were identified by HPLC. In the reaction mixture, AMP (3.5 min), nicotinamide (Nic 5.5 min), and NAD (6.5 min) were observed. An additional component to the reaction mixture (labelled X) was observed (6.1 min), but could not be further identified by mass spectrometry. Data in all panels are one representative from three independent experiments with similar results.



**Extended Data Figure 10 | Detection of the ubiquitination intermediate by using SdeA$_{519-1100}$.** **a**, Full-length SdeA cannot produce $^{32}$P-labelled product in reactions using $^{32}$P-α-NAD. Reaction samples resolved by SDS–PAGE were detected by Coomassie staining (left panel) and then by autoradiography (right panel). Note the $^{32}$P-α-AMP-GST-ubiquitin complex can be detected in the reaction containing E1 but not SdeA. **b**, **c**, SdeA$_{519-1100}$ is defective in auto-ubiquitination. Reactions containing the indicated components were allowed to proceed for the indicated time duration and the production of ubiquitinated Rab33b (**b**) or SdeA$_{519-1100}$ was detected by immunoblotting. **d**, SdeA$_{519-1100}$ induces the production of nicotinamide from NAD and ubiquitin. Retention time for nicotinamide and NAD was first determined by HPLC and nicotinamide can only be detected in the reaction containing SdeA$_{519-1100}$, NAD and ubiquitin. **e**, SdeA$_{519-1100}$ induces the production of $^{32}$P-ADPR-labelled ubiquitin. GST-ubiquitin or GST–ubiquitin$_{R42A}$ was incubated with $^{32}$P-α-NAD and SdeA$_{519-1100}$ for 0036 h. Classical E1 incubated with GST–ubiquitin was included as a control. Samples resolved by SDS–PAGE before autoradiography (20 min) (right panel). Note that GST–ubiquitin$_{R42A}$ cannot be labelled by $^{32}$P. Data in panels **a**–**e** are one representative from two independent experiments with similar results. **f**, The detection of a peptide with *m*/*z* 737.33 corresponding to the tryptic peptide E$_{34}$GIPPDQQRLIFAGK$_{48}$ containing one ADP-ribosylation site was detected only after ubiquitin was incubated with SdeA$_{519-1100}$. As a loading control, another unmodified ubiquitin peptide T$_{55}$LSDYNIQK$_{63}$ was detected in both control and treated samples. **g**, Tandem mass analysis revealed that ADP-ribosylation occurred on Arg42 evidenced by the extensive fragmentation of the ADP-ribosylation into adenine, adenosine, AMP and ADP ions. Although not as extensive, the fragmentation of the peptide backbone helps confirm the peptide sequence. Data shown in all panels are one representative from two independent experiments with similar results. **a**–**c**, **e**, Uncropped blots and autoradiograph images are shown in Supplementary Fig. 1.

doi:10.1038/nature17664

# Efficient introduction of specific homozygous and heterozygous mutations using CRISPR/Cas9

Dominik Paquet[1]*, Dylan Kwart[1]*, Antonia Chen[1], Andrew Sproul[2]†, Samson Jacob[2], Shaun Teo[1], Kimberly Moore Olsen[1], Andrew Gregg[1,3], Scott Noggle[2] & Marc Tessier-Lavigne[1]

**The bacterial CRISPR/Cas9 system allows sequence-specific gene editing in many organisms and holds promise as a tool to generate models of human diseases, for example, in human pluripotent stem cells[1,2]. CRISPR/Cas9 introduces targeted double-stranded breaks (DSBs) with high efficiency, which are typically repaired by non-homologous end-joining (NHEJ) resulting in nonspecific insertions, deletions or other mutations (indels)[2]. DSBs may also be repaired by homology-directed repair (HDR)[1,2] using a DNA repair template, such as an introduced single-stranded oligo DNA nucleotide (ssODN), allowing knock-in of specific mutations[3]. Although CRISPR/Cas9 is used extensively to engineer gene knockouts through NHEJ, editing by HDR remains inefficient[3–8] and can be corrupted by additional indels[9], preventing its widespread use for modelling genetic disorders through introducing disease-associated mutations. Furthermore, targeted mutational knock-in at single alleles to model diseases caused by heterozygous mutations has not been reported. Here we describe a CRISPR/Cas9-based genome-editing framework that allows selective introduction of mono- and bi-allelic sequence changes with high efficiency and accuracy. We show that HDR accuracy is increased dramatically by incorporating silent CRISPR/Cas-blocking mutations along with pathogenic mutations, and establish a method termed 'CORRECT' for scarless genome editing. By characterizing and exploiting a stereotyped inverse relationship between a mutation's incorporation rate and its distance to the DSB, we achieve predictable control of zygosity. Homozygous introduction requires a guide RNA targeting close to the intended mutation, whereas heterozygous introduction can be accomplished by distance-dependent suboptimal mutation incorporation or by use of mixed repair templates. Using this approach, we generated human induced pluripotent stem cells with heterozygous and homozygous dominant early onset Alzheimer's disease-causing mutations in amyloid precursor protein (APP$^{\mathrm{Swe}}$)[10] and presenilin 1 (PSEN1$^{\mathrm{M146V}}$)[11] and derived cortical neurons, which displayed genotype-dependent disease-associated phenotypes. Our findings enable efficient introduction of specific sequence changes with CRISPR/Cas9, facilitating study of human disease.**

While attempting to knock-in early onset Alzheimer's disease mutations into iPS cells using CRISPR/Cas9, we detected HDR by presence of an intended mutation provided via the cognate ssODNs, however most HDR events also contained unwanted indels (Fig. 1a). This is presumably due to the high nuclease activity of CRISPR/Cas9 (refs 3, 4, 6, 8), which may continuously re-cut edited loci until sufficient modification by NHEJ prevents further targeting. If so, this re-editing may be blocked by simultaneously mutating the NGG protospacer adjacent motif (PAM) or guide RNA binding sequence, which CRISPR/Cas9 requires for targeting[2], as shown in prokaryotes[12]. As the efficacy of potential blocking mutations has not been systematically studied in eukaryotic cells, we tested their effect on HDR accuracy in wild-type human induced pluripotent stem cells (iPS cells) (Extended Data Fig. 1) and, for comparison, human embryonic kidney (HEK293) cells. We introduced Cas9–eGFP and single guide RNA (sgRNA) plasmids together with five pooled repair ssODN templates, which in addition to the APP$^{\mathrm{Swe}}$ or PSEN1$^{\mathrm{M146V}}$ pathogenic mutation also contained a putative silent CRISPR/Cas-blocking mutation in the PAM or guide RNA target sequence, or a control non-blocking mutation outside those regions (Fig. 1b, c and Supplementary Tables 1 and 2).

We analysed genomic loci of Cas9–eGFP-expressing cells by next-generation sequencing and determined the fraction of HDR reads that were 'accurate', that is, without undesirable indel modifications (Fig. 1d, e; Extended Data Table 1 lists overall HDR rates for all experiments). Without blocking mutations, only 6 to 35% of reads that incorporated pathogenic mutations had accurate HDR, but presence of a CRISPR/Cas-blocking PAM mutation increased HDR accuracy in both iPS cells and HEK293 cells two- to tenfold, depending on locus and cell type, which may increase the probability of accurately editing both alleles in a cell up to 100-fold (assuming independent allele editing). The remaining 'inaccurate' HDR events were presumably generated by prior or concomitant NHEJ. Blocking mutations targeting the guide RNA sequence increased HDR accuracy to a similar extent for *APP*, but much less for *PSEN1* (Fig. 1d, e). Therefore, whereas PAM-site mutations seem broadly effective, guide RNA target mutations may have variable effects at different loci. Similar results were obtained for ssODNs transfected individually rather than pooled (Extended Data Fig. 2a, b). Indel frequency, position, and size had expected distributions[1,13] (Extended Data Fig. 3a). Interestingly, in experiments with pooled ssODNs, up to 11% of HDR reads contained multiple blocking or control mutations (Extended Data Fig. 2c, d), showing that cells used multiple oligonucleotides in multiple rounds of repair, and highlighting the propensity of CRISPR/Cas9 for re-editing. Thus, CRISPR/Cas-blocking mutations, preferably in the PAM, minimize undesirable re-editing during derivation of knock-in mutant clones.

Introducing silent blocking mutations in coding regions is often possible, though in some cases silent mutations may be precluded by the PAM reading frame or prove ineffective in the guide RNA target. Furthermore, in non-coding regions, blocking mutations may have unwanted consequences. Intended mutations may occasionally double as blocking mutations, but this is not always the case. We therefore developed a method to remove blocking mutations when desired, termed CORRECT (consecutive re-guide or re-Cas steps to erase CRISPR/Cas-blocked targets), with two variants: re-guide and re-Cas. In both, blocking mutations are first introduced together with

[1]Laboratory of Brain Development and Repair, The Rockefeller University, 1230 York Avenue, New York, New York 10065, USA. [2]The New York Stem Cell Foundation Research Institute, New York, New York 10032, USA. [3]Weill Cornell Graduate School of Medical Sciences, The Rockefeller University and Sloan-Kettering Institute Tri-institutional MD-PhD Program, 1300 York Avenue, New York, New York 10065, USA. †Present address: Department of Pathology and Cell Biology and the Taub Institute for Research on Alzheimer's Disease and the Aging Brain, Columbia University Medical Center, 630 West 168th Street, New York, New York 10032, USA.
*These authors contributed equally to this work.

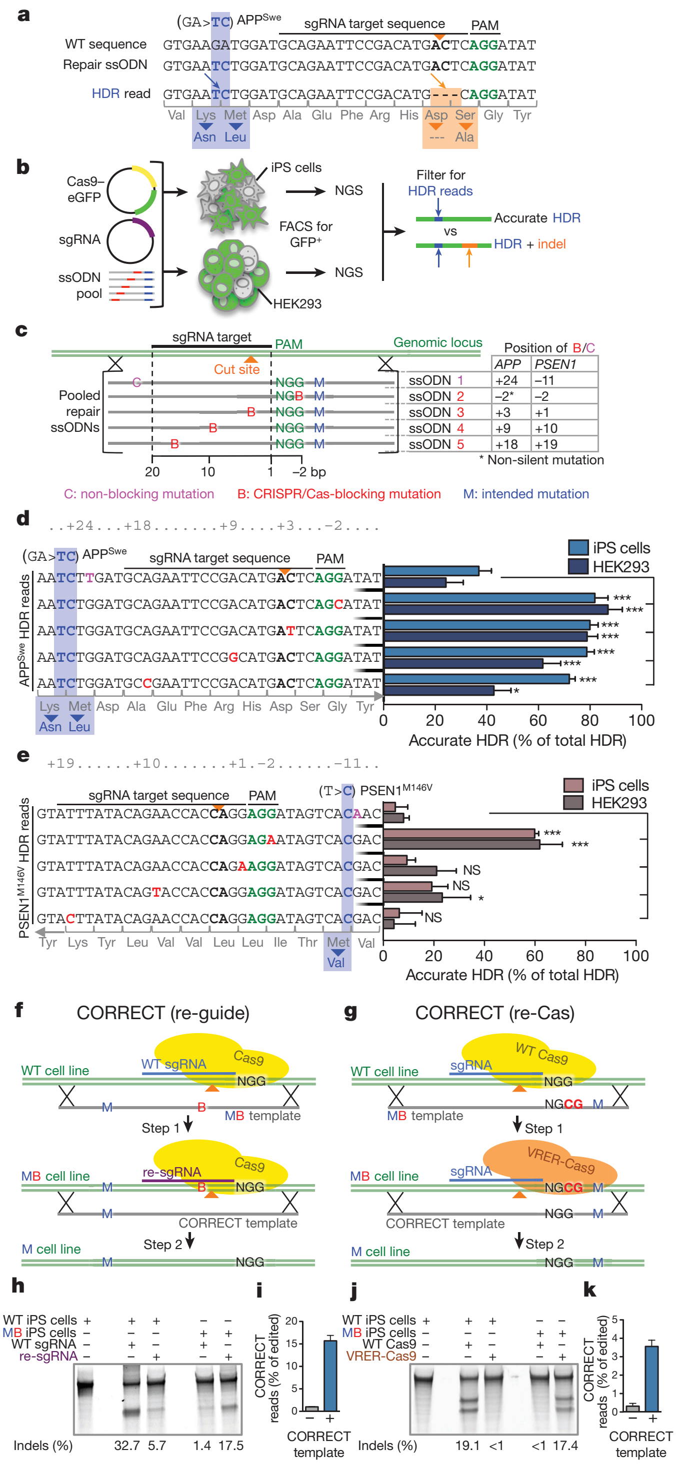


**Figure 1 | CRISPR/Cas-blocking mutations increase HDR accuracy by preventing re-editing and can be used for scarless CORRECT editing.** **a**, *APP* sequencing alignment showing concomitant HDR (blue arrow) and indels (orange arrow) after editing. **b**, Experimental setup for gene editing analysis by next-generation sequencing (NGS). **c**, Pooled ssODNs used to test effects of CRISPR/Cas-blocking mutations. **d**, **e**, Percentages of accurate HDR for blocking or control mutations at *APP* (**d**) and *PSEN1* (**e**) loci in iPS cells and HEK293 cells. Values represent mean ± s.e.m. ($n = 3$). ***$P < 0.001$, **$P < 0.01$, *$P < 0.05$, one-way ANOVA. **f**, **g**, Two-step workflow for CORRECT variants re-guide (**f**) and re-Cas (**g**): re-guide uses a blocking mutation B in the guide RNA target sequence, whereas for re-Cas the PAM is mutated to a sequence detected by a Cas9 variant. Blocking mutations are removed in step 2 using re-sgRNA/WT-Cas9 or WT-sgRNA/VRER-Cas9, while pathogenic mutations M are retained. **h**, **j**, Surveyor mismatch cleavage assay detecting CRISPR/Cas9 activity shows specificity of WT-Cas9/WT-sgRNAs for wild-type targets, and WT-Cas9/re-sgRNA (**h**) or VRER-Cas9/WT-sgRNA (**j**) for mutated loci. **i**, **k**, Next-generation sequencing quantification of genomes with sequence inserted by HDR with CORRECT templates in pooled iPS cells ($n = 2$). WT, wild-type.

intended mutations by HDR. Whereas the re-guide blocking mutation interferes with guide RNA targeting, the re-Cas mutation blocks PAM detection by mutating the NGG to the target sequence of a specificity-modified Cas9 (in our experiments, NGCG, target of the recently described VRER-Cas9 (ref. 14)). The blocking mutation is then removed with modified reagents: for re-guide, a re-sgRNA targeting the modified sequence is used with wild-type Cas9 (Fig. 1f). For re-Cas, the modified PAM is targeted with the Cas9 variant (Fig. 1g). We tested the feasibility of CORRECT by re-guide using an $APP^{Swe}$ iPS cell line containing a guide RNA target mutation (see Fig. 1); for re-Cas, we generated an $APP^{A673T}$ mutant iPS cell line with a NGCG PAM mutation (Extended Data Fig. 2e, f). We then removed the blocking mutations from both lines with CORRECT templates and wild-type Cas9/re-sgRNA (for $APP^{Swe}$) or VRER-Cas9/sgRNA (for $APP^{A673T}$) (Fig. 1h, j). The expected editing events were detected with high efficiency by next-generation sequencing (Fig. 1i, k). Thus, CORRECT enables efficient scarless introduction of just an intended mutation.

We next examined mutational status of the two alleles in individual iPS cell clones. We could readily isolate clones with homozygous early onset Alzheimer's disease mutations, but, interestingly, in clones heterozygous for early onset Alzheimer's disease mutations, the 'non-HDR' allele almost always contained indels (Extended Data Figs 3b and 4a, b). This is possibly due to the high efficiency of Cas9 (refs 3, 4, 6, 8), which results mostly in bi-allelic modifications[4–7], and raises the question of how to isolate heterozygous clones.

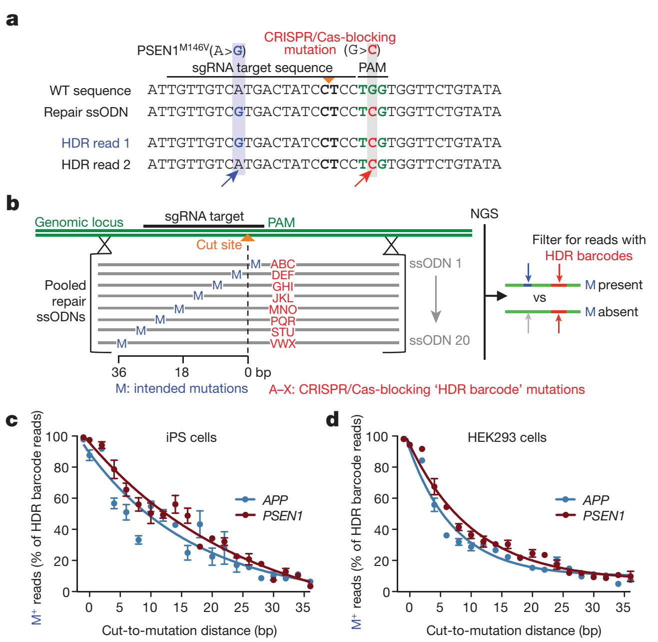


**Figure 2 | A monotonic inverse relationship between mutation incorporation and distance from the CRISPR/Cas9 cleavage site.** **a**, *PSEN1* sequencing alignment showing introduction of a CRISPR/Cas-blocking mutation (red arrow) with or without the pathogenic mutation (blue arrow) during HDR. **b**, Pooled ssODNs used to scan mutation incorporation rates based on cut-to-mutation distance. Barcode mutations (red) identify HDR-reads and mutation M position during next-generation sequencing (NGS) analysis. **c**, **d**, A monotonic inverse relationship governs rate of mutation M incorporation and cut-to-mutation distance during HDR in both iPS cells (**c**) and HEK293 cells (**d**) ($n = 4$ for iPS, $n = 3$ for HEK293); goodness of fit: $R^2$ (*APP*) = 0.75 (iPS) / 0.96 (HEK293), $R^2$ (*PSEN1*) = 0.94 (iPS) / 0.97 (HEK293); curves for *APP* and *PSEN1* are not significantly different, two-tailed *t*-test: $P = 0.31$ (iPS) / 0.06 (HEK293).

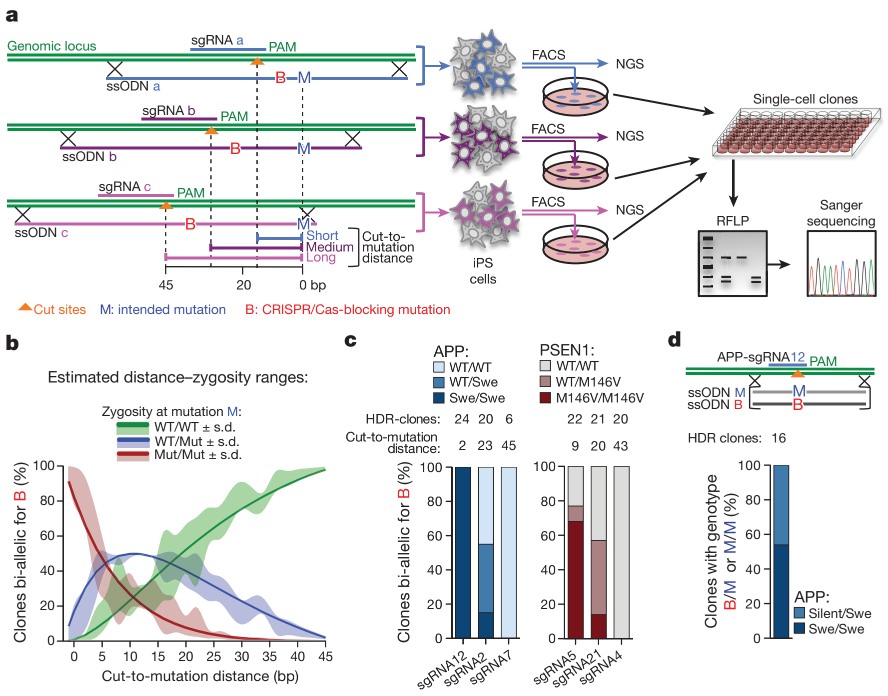



**Figure 3 | Introduction of heterozygous or homozygous mutations into iPS cells by manipulating the cut-to-mutation distance or using mixed HDR templates. a**, Experimental setup with three sgRNA/ssODN pairs per locus with increasing cut-to-mutation distance. Edited iPS cells were analysed by next-generation sequencing (NGS; see Extended Data Fig. 5) or grown for clonal analysis. **b**, Predicted distance ranges for desired zygosities, calculated based on oligonucleotide scan data (see Fig. 2c and Methods). **c**, Frequency of different *APP* and *PSEN1* mutation genotypes in single-cell clones with bi-allelic HDR of blocking mutations. Indicated zygosities fit to predicted values. **d**, Introduction of heterozygous mutations with mixed repair ssODNs. One ssODN contained the pathogenic $APP^{Swe}$ mutation, which blocks sgRNA12 (M), the other a silent blocking mutation (B) (see alternative approach with two silent blocking mutations in Extended Data Fig. 5d).

A first approach came from our observation that many alleles that incorporated a silent CRISPR/Cas-blocking mutation did not contain the intended pathogenic mutation, particularly if it was distant from the CRISPR/Cas9 cleavage site (Fig. 2a). This was similar to reports of distance dependence for editing with CRISPR/Cas9 (refs 9, 15) or other systems[13,16–19]. We reasoned that a predictable relationship between distance and mutation incorporation could be exploited to control allelic mutation incorporation. We therefore characterized distance dependence at the *APP* and *PSEN1* loci by scanning mutation incorporation rates with 20 different pooled ssODNs, each containing a unique CRISPR/Cas-blocking three-base-pair barcode sequence, as well as single point mutations at increasing distances from the cleavage site (Fig. 2b and Supplementary Table 2). Notably, we found a clear monotonic inverse relationship between rate of mutation incorporation and distance from cleavage site that did not differ significantly for *APP* and *PSEN1* in either iPS cells or HEK293 cells (Fig. 2c, d). The relationship was also similar for longer ssDNA or dsDNA HDR repair templates (Extended Data Fig. 4d), and for three distinct sgRNA/ssODN pairs (Supplementary Tables 1 and 2) targeting DSBs at short, intermediate and long cut-to-mutation distances (Fig. 3a and Extended Data Fig. 5a, b). Thus, a general and predictable 'distance effect' may govern mutation incorporation by HDR during gene editing in these human cells.

Our data imply that cut-to-mutation distance needs to be minimized for efficient homozygous mutation incorporation and, conversely, that frequencies of mono-allelic alterations should increase at greater distances, as mutation incorporation probability drops. We determined overall probability of mutation incorporation for iPS cells by combining $APP^{Swe}$ and $PSEN1^{M146V}$ oligonucleotide scan data (from Fig. 2c) and calculated expected distance ranges favouring homozygous, heterozygous and wild-type genotypes by multiplying single allele probabilities (assuming independent editing at both alleles) (Fig. 3b). To test these predictions, we derived single-cell clones from iPS cells electroporated with the abovementioned sgRNA/ssODN pairs (Fig. 3a), and selected those with bi-allelic incorporation of silent CRISPR/Cas-blocking mutations (Fig. 3c). The rate of homozygosity and heterozygosity for the pathogenic mutation correlated with our predictions, indicating that cut-to-mutation distance can be exploited to control zygosity using Fig. 3b to select distance.

At certain loci, only guide RNAs targeting close to the intended mutation may be available, which could preclude isolation of heterozygous clones using the distance effect. As an alternative, we considered equimolar mixing of two ssODNs that both possess a blocking mutation, but only one of which contains the pathogenic mutation (Fig. 3d, Extended Data Fig. 5c; alternative approach in Extended Data Fig. 5d). We validated this approach using the closely targeting APP-sgRNA12, which previously only yielded clones homozygous for the $APP^{Swe}$ mutation, and detected many with mono-allelic incorporation. We also verified this strategy for $PSEN1^{M146V}$ (Extended Data Fig. 5e), suggesting it is widely applicable.

Many genetic disorders have been studied by deriving iPS cells from patients with a disease, but this approach takes several months, and is limited by availability of patient cells and variable genetic backgrounds. These problems can be circumvented by knock-in of disease mutations in a reference 'wild-type' cell line, which only takes a few weeks and provides isogenic controls. Alzheimer's disease has mostly been studied in animal models relying on non-physiological mutant gene overexpression[20]. Human iPS cells derived from patients with early onset Alzheimer's disease mutations were recently established[21–25], but only TALEN-mediated gene editing has been used to knock-in an early onset Alzheimer's disease mutation[26]. Using distance to control zygosity, we generated allelic series with knock-in mutations in $APP^{Swe}$ or $PSEN1^{M146V}$ (Extended Data Fig. 6a, b and Extended Data Table 2). We differentiated iPS cells into cortical neurons (Extended Data Fig. 6c–j), and examined whether $APP^{Swe}$ and $PSEN1^{M146V}$ mutations increase total amyloid-β (Aβ) generation or the ratio of the 42-residue

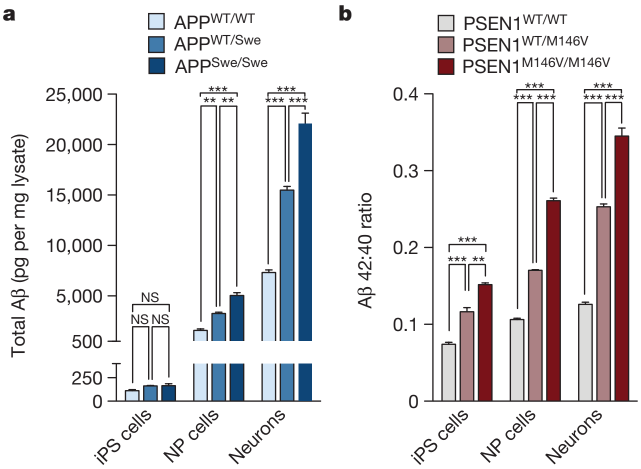



**Figure 4 | $APP^{Swe}$ and $PSEN1^{M146V}$ knock-in lines display genotype-dependent disease-associated changes in Aβ secretion. a**, Mutation load dependent changes in total Aβ in $APP^{Swe}$ mutant iPS cells, neural precursors and cortical neurons. NP, neural precursor. **b**, Mutation load dependent changes in Aβ42:40 ratios in $PSEN1^{M146V}$ mutant cells. Values represent mean ($n=3$ biological replicates) ± s.e.m. $^{**}P<0.05$ and $^{***}P<0.001$, one-way ANOVA.

versus the 40-residue Aβ peptide (Aβ42:40), respectively, as predicted from patient data and model systems[10,11]. We found more than three-fold higher Aβ levels in homozygous, and twofold higher Aβ levels in heterozygous $APP^{Swe}$ mutant cells, and up to threefold increase in secreted Aβ42:40 ratio in homozygous and twofold increase in heterozygous $PSEN1^{M146V}$ mutant cells, compared to isogenic controls (Fig. 4a, b). Changes in Aβ levels and Aβ42:40 ratios correlated with neuronal identity and maturity (Fig. 4a, b). Thus Alzheimer's disease related phenotypes can be faithfully modelled in human neurons by introducing early onset Alzheimer's disease associated mutations, and these phenotypes correlate with mutation load.

Widespread application of CRISPR/Cas9 to induce specific genomic changes depends on strategies to improve accurate HDR. Manipulations of cell cycle and small molecules inhibiting NHEJ have recently been reported to increase HDR rates[27–30], but these approaches do not directly aim to improve HDR accuracy, achieved here using CRISPR/Cas-blocking mutations. This allowed us to isolate one accurately edited line by picking just 20 to 40 clones on average (Extended Data Table 1), a rate compatible with manual picking, which might be further improved by combination with small molecule inhibitors of NHEJ. Titrating down Cas9 or guide RNA levels may also improve accuracy, but in our experiments this greatly reduced HDR rates such that manual single-cell clone picking became impractical (data not shown). Methods improving rate and accuracy of HDR can also be combined with CORRECT, enabling efficient scarless editing in dividing cells.

To enable control of zygosity during CRISPR/Cas9 editing, we extended previous studies[9,13,15–19] by characterizing in two human cell types the stereotyped inverse relationship between incorporation rate of a base by HDR and its distance from CRISPR/Cas9 cleavage site. The length of gene conversion tracts we observed for CRISPR/Cas9 editing (~30–35 bp) was similar to that for TALENs in human cells[13,18], but differed markedly for zinc finger nucleases in *Drosophila* (over 3,000 bp)[17] and restriction enzymes in rodent cells (80–200 bp)[16,19], potentially reflecting experimental or species differences (for example, in activities of repair pathways). Controlling zygosity by exploiting the distance effect may work best in systems with short gene conversion tracts. Our alternative approach of oligonucleotide mixing is more universally applicable.

The distance relationship did not change with altered editing conditions including HDR template types and therefore probably reflects intrinsic features of the repair mechanism. Distance dependence may reflect the distribution of different size deletions after CRISPR/Cas9-mediated DSBs, which require only the part of the ssODN overlapping the deletion for repair[16] (see model in Extended Data Fig. 7). Regardless of mechanism, the observation of a stereotyped distance effect implies that HDR is most efficiently achieved by selecting guide RNAs targeting close to the intended sequence change, and allows definition of optimal distance ranges for improved guide RNA selection to generate mono- or bi-allelic modifications.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 26 May 2015; accepted 11 March 2016.**
**Published online 27 April 2016.**

1. Mali, P. *et al.* RNA-guided human genome engineering via Cas9. *Science* **339,** 823–1278 (2013).
2. Hsu, P. D., Lander, E. S. & Zhang, F. Development and applications of CRISPR-Cas9 for genome engineering. *Cell* **157,** 1262–1278 (2014).
3. Cong, L. *et al.* Multiplex genome engineering using CRISPR/Cas systems. *Science* **339,** 819–823 (2013).
4. Dow, L. E. *et al.* Inducible *in vivo* genome editing with CRISPR-Cas9. *Nature Biotechnol.* **33,** 390–394 (2015).
5. Platt, R. J. *et al.* CRISPR-Cas9 knockin mice for genome editing and cancer modeling. *Cell* **159,** 440–455 (2014).
6. Wang, H. *et al.* One-step generation of mice carrying mutations in multiple genes by CRISPR/Cas-mediated genome engineering. *Cell* **153,** 910–918 (2013).
7. Canver, M. C. *et al.* Characterization of genomic deletion efficiency mediated by clustered regularly interspaced palindromic repeats (CRISPR)/Cas9 nuclease system in mammalian cells. *J. Biol. Chem.* **289,** 21312–21324 (2014).
8. Yang, H. *et al.* One-step generation of mice carrying reporter and conditional alleles by CRISPR/Cas-mediated genome engineering. *Cell* **154,** 1370–1379 (2013).
9. Inui, M. *et al.* Rapid generation of mouse models with defined point mutations by the CRISPR/Cas9 system. *Sci. Rep.* **4,** 5396 (2014).
10. Haass, C. *et al.* The Swedish mutation causes early-onset Alzheimer's disease by beta-secretase cleavage within the secretory pathway. *Nature Med.* **1,** 1291–1296 (1995).
11. Alzheimer's Disease Collaborative Group. The structure of the presenilin 1 (*S182*) gene and identification of six novel mutations in early onset AD families. *Nature Genet.* **11,** 219–222 (1995).
12. Jiang, W., Bikard, D., Cox, D., Zhang, F. & Marraffini, L. A. RNA-guided editing of bacterial genomes using CRISPR-Cas systems. *Nature Biotechnol.* **31,** 233–239 (2013).
13. Yang, L. *et al.* Optimization of scarless human stem cell genome editing. *Nucleic Acids Res.* **41,** 9049–9061 (2013).
14. Kleinstiver, B. P. *et al.* Engineered CRISPR-Cas9 nucleases with altered PAM specificities. *Nature* **523,** 481–485 (2015).
15. Bialk, P., Rivera-Torres, N., Strouse, B. & Kmiec, E. B. Regulation of gene editing activity directed by single-stranded oligonucleotides and CRISPR/Cas9 systems. *PLoS ONE* **10,** e0129308 (2015).
16. Elliott, B., Richardson, C., Winderbaum, J., Nickoloff, J. A. & Jasin, M. Gene conversion tracts from double-strand break repair in mammalian cells. *Mol. Cell. Biol.* **18,** 93–101 (1998).
17. Beumer, K. J., Trautman, J. K., Mukherjee, K. & Carroll, D. Donor DNA utilization during gene targeting with zinc-finger nucleases. *Genes Genomes Genetics (G3)* **3,** 657–664 (2013).
18. Rivera-Torres, N., Strouse, B., Bialk, P., Niamat, R. A. & Kmiec, E. B. The position of DNA cleavage by TALENs and cell synchronization influences the frequency of gene editing directed by single-stranded oligonucleotides. *PLoS ONE* **9,** e96483 (2014).
19. Taghian, D. G. & Nickoloff, J. A. Chromosomal double-strand breaks induce gene conversion at high frequency in mammalian cells. *Mol. Cell. Biol.* **17,** 6386–6393 (1997).
20. Götz, J. & Ittner, L. M. Animal models of Alzheimer's disease and frontotemporal dementia. *Nature Rev. Neurosci.* **9,** 532–544 (2008).
21. Yagi, T. *et al.* Modeling familial Alzheimer's disease with induced pluripotent stem cells. *Hum. Mol. Genet.* **20,** 4530–4539 (2011).
22. Israel, M. A. *et al.* Probing sporadic and familial Alzheimer's disease using induced pluripotent stem cells. *Nature* **482,** 216–220 (2012).
23. Kondo, T. *et al.* Modeling Alzheimer's disease with iPSCs reveals stress phenotypes associated with intracellular Aβ and differential drug responsiveness. *Cell Stem Cell* **12,** 487–496 (2013).
24. Muratore, C. R. *et al.* The familial Alzheimer's disease APPV717I mutation alters APP processing and Tau expression in iPSC-derived neurons. *Hum. Mol. Genet.* **23,** 3523–3536 (2014).
25. Sproul, A. A. *et al.* Characterization and molecular profiling of PSEN1 familial Alzheimer's disease iPSC-derived neural progenitors. *PLoS ONE* **9,** e84547 (2014).
26. Woodruff, G. *et al.* The presenilin-1 ΔE9 mutation results in reduced γ-secretase activity, but not total loss of PS1 function, in isogenic human stem cells. *Cell Rep.* **5,** 974–985 (2013).

27. Lin, S., Staahl, B. T., Alla, R. K. & Doudna, J. A. Enhanced homology-directed human genome engineering by controlled timing of CRISPR/Cas9 delivery. *eLife* **3,** e04766 (2015).
28. Yu, C. *et al.* Small molecules enhance CRISPR genome editing in pluripotent stem cells. *Cell Stem Cell* **16,** 142–147 (2015).
29. Maruyama, T. *et al.* Increasing the efficiency of precise genome editing with CRISPR-Cas9 by inhibition of nonhomologous end joining. *Nature Biotechnol.* **33,** 538–542 (2015).
30. Chu, V. T. *et al.* Increasing the efficiency of homology-directed repair for CRISPR-Cas9-induced precise gene editing in mammalian cells. *Nature Biotechnol.* **33,** 543–548 (2015).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** This research was supported by The Rockefeller University, The New York Stem Cell Foundation, The Ellison Foundation, Cure Alzheimer's Fund, the Empire State Stem Cell fund through New York State Department of Health contract number C023046, and CTSA, RUCCTS grant number 8 UL1 TR000043 from the National Center for Advancing Translational Sciences (NCATS, NIH). D.P. is a New York Stem Cell Foundation Druckenmiller Fellow and received a fellowship from the German Academy of Sciences Leopoldina. D.K. is a Howard Hughes Medical Institute International Student Research Fellow and received a fellowship from the National Sciences and Engineering Research Council of Canada. S.T. is supported by the Agency for Science, Technology and Research of Singapore. A.G. is supported by a Medical Scientist Training Program grant from the National Institute of General Medical Sciences of the National Institutes of Health under award number T32GM007739 to the Weill Cornell/Rockefeller/Sloan-Kettering Tri-institutional MD-PhD program. We thank members of the Tessier-Lavigne laboratory and L. Marraffini for discussions. Our thanks also go to S. Mazel and the team at the Rockefeller Univeristy Flow Cytometry Resource Center, J. Gonzalez and the team at the Rockefeller University Translational Technology Core Laboratory, C. Zhao and the team at the Rockefeller University Genomics Resource Center, and D. Paull and M. Duffield for technical help. Opinions expressed here are solely those of the authors and do not necessarily reflect those of the Empire State Stem Cell Fund, the New York State Department of Health, or the State of New York. The content of this study is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

**Author Contributions** D.P., D.K. and M.T.-L. conceived and designed the study. D.P. and D.K. performed and analysed the experiments. A.C. and A.G. helped perform the experiments. S.T. helped analyse next-generation sequencing data. A.S., S.J. and S.N. generated and characterized the iPS cells. K.M.O. performed and analysed the electrophysiology assays. D.P., D.K., and M.T.-L. wrote the manuscript with input from all authors.

**Author Information** Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.T.-L. (marctl@rockefeller.edu).

## METHODS

**sgRNA and Cas9-VRER plasmid design and construction.** sgRNAs were designed using the Zhang laboratory CRISPR design tool (http://crispr.mit.edu). sgRNA sequences targeting *APP* or *PSEN1* (Supplementary Table 1) were cloned into plasmid MLM3636 (a gift from K. Joung, Addgene number 43860) as previously described[31]. To generate the Cas9-VRER variant[14] with human codon usage, we introduced the 4 mutations into pCas9_GFP (a gift from K. Musunuru, Addgene plasmid number 44719). Briefly, we amplified fragments around the intended mutation sites by PCR with mutated primers (Supplementary Table 1), digested the plasmid with BamHI/BsrGI and fused all fragments by Gibson assembly.

**Design of ssODN repair templates.** The 100-nt ssODN repair templates (PAGE-purified, IDT) were designed with homologous genomic flanking sequence centred around the predicted CRISPR/Cas9 cleavage site and containing pathogenic and/or CRISPR/Cas-blocking mutations (Supplementary Table 2). CRISPR/Cas-blocking silent (that is, that do not alter the amino acid sequence) mutations were selected based on codon-usage of the edited gene by changing the codon to another codon already used in the same mRNA for the respective amino acid.

**Generation of long ssDNA and dsDNA repair templates.** To generate 200 bp and 400 bp ssDNA and dsDNA repair templates, 1,000 bp of *PSEN1* sequence around the edited locus was first PCR-amplified and TOPO-cloned. Then, a library of 20 ssODN oligonucleotides or gBlocks (IDT) containing the required mutations was integrated into the TOPO-vector by Gibson assembly (NEB), resulting in a library of 20 plasmid templates, each containing CRISPR/Cas-blocking barcode mutations and an intended mutation at varying cut-to-mutation distances (as described in Fig. 2b). From each plasmid template, 200 bp and 400 bp dsDNA PCR amplicons were generated (primers in Supplementary Table 1), and mixed in equal amounts to generate pools of either size PCR template amplicons. Template pools were then gel extracted to remove residual plasmid. These were then re-amplified by PCR and concentrated before transfection. To generate ssDNA templates, dsDNA amplicons were generated as described above with 5′ phosphorylated forward primers. Re-amplified dsDNA amplicons were then digested with lambda exonuclease (NEB) to generate ssDNA. Reactions were column purified before transfection (see Extended Data Fig. 4c).

**Immunocytochemistry and microscopy.** Cells were fixed in 4% paraformaldehyde, permeabilized in PBS/0.1% Triton X-100 and stained with primary and secondary antibodies (see later). Stained cells were imaged on a Nikon Eclipse Ti inverted microscope and acquired using NIS Elements imaging software (Nikon). Fiji (http://www.Fiji.sc) and Adobe Photoshop were used to pseudo-colour images, adjust contrast and add scale bars.

**Antibodies.** The following antibodies were used: Oct4 (1:500, Stemgent S090023), Tra160 (1:500, Millipore MAB4360), SSEA4 (1:500, Abcam ab16287), Nanog (1:500, Cell Signaling 4903), MAP2 (1:2000, Abcam 5392), Pax6 (1:300, Covance PRB-278P), Tuj1 (mouse 1:1,000, Covance MMS-435P / rabbit 1:1,000, Covance MRB-435P), Otx2 (1:100, Millipore AB9566), Nestin (1:200, Millipore 2C13B9), FoxG1 (1:300, Abcam ab18259), CTIP2 (1:300, Abcam ab18465), Tbr1 (1:500, Millipore AB2261), SatB2 (1:100, Abcam ab51502), MAGUK (1:100, NeuroMab K28_86), Synapsin (1:200, Cell Signalling Technologies 5297), anti-mouse/rabbit/rat/chicken Alexa Fluor 488/568/647 (Invitrogen 1:500).

**iPS cell lines.** iPS cells were reprogrammed from human skin fibroblasts (Coriell Institute, catalog ID: AG07889) of a 18-year-old male individual using the Cytotune-iPS Sendai Reprogramming Kit (Life Technologies) according to the manufacturer's instructions, following Rockefeller University Institutional Review Board approval. Informed consent was obtained from all subjects upon sample submission to Coriell Institute. Fibroblasts were confirmed to be wild-type for all studied loci by genotyping. Multiple clones were selected based on characteristic morphology. Genetic fingerprinting confirmed iPS cells were derived from corresponding fibroblast lines. Clone 7889SA possessed a normal karyotype (Cell Line Genetics), and was characterized for typical iPS cell properties and absence of mycoplasma contamination.

Expression of pluripotency genes was analysed by NanoString nCounter gene expression system using a pre-designed codeset[32]. Data was normalized to the geometric mean of three housekeeping genes (ACTB, POLR2A, ALAS1) using the nSolver Analysis Software v1.0 (NanoString). 100 ng of total RNA from line 7889SA was compared to RNA extracted from the human embryonic stem cell lines HUES9 (ref. 33). Gene expression for 7 pluripotency markers and the four Yamanaka factors (Oct4, Sox2, Klf4, c-Myc) introduced as Sendai transgenes (s-t) was compared. Note that the s-tSox2 probe detects some expression of endogenous Sox2, leading to larger values for both lines.

Expression of pluripotency markers Oct4, Tra160, SSEA4 and Nanog was confirmed by immunofluorescence. *In vivo* pluripotency was confirmed by teratoma analysis as described[25,32]. Briefly, undifferentiated iPS cells were embedded into Matrigel and subcutaneously injected into the dorsal flank of two immune-compromised three-month-old male or female mice (NOD.Cg-*Prkdc$^{scid}$Il2rg$^{tm1Wjl}$*/SzJ, stock no. 005557, The Jackson Laboratory). Paraffin sections of the teratomas were subjected to haematoxylin and eosin (H&E) staining and structures characteristic for the three germ layers (ectoderm, mesoderm and endoderm) were identified by microscopy. Animal work was approved by the Columbia Institutional Animal Care and Use Committee and no randomization or blinding was used for analysis.

To generate homozygous and heterozygous APP$^{Swe}$ iPS cell lines, cells were electroporated with the sgRNA2/ssODN and sgRNA12/ssODN combinations described in Fig. 3c and Supplementary Table 2. To study heterozygous and homozygous PSEN1$^{M146V}$ mutations, cells were electroporated with the sgRNA5/ssODN combination described in Fig. 3c and Supplementary Table 2. Electroporated cells were isolated by FACS, followed by single-cell clone generation, RFLP and sequencing analysis as described below. One iPS cell line per genotype was isolated and characterized. The newly established gene-edited lines displayed normal karyotypes and expressed pluripotency markers Oct4, Tra160, SSEA4, Nanog and alkaline phosphatase (data not shown).

**Cell culture and transfection.** iPS cells were maintained on irradiated MEFs (Globalstem) plated on cell culture plates coated with 0.1% gelatin and grown in HUESM (Knockout Dulbecco's modified Eagle's Medium (KO-DMEM), 20% knockout serum, 0.1 mM non-essential amino acids, 2 mM Glutamax, 100 U per ml penicillin, 0.1 mg per ml streptomycin (all Life Technologies), 0.1 mM 2-mercaptoethanol (Sigma-Aldrich), 10 ng ml$^{-1}$ FGF2 (Stemgent), at 37 °C with 5% $CO_2$. Prior to transfection, iPS cells were transferred to Geltrex-coated (Life Technologies) cell culture plates and grown in MEF-conditioned HUESM containing 10 μM ROCK inhibitor (Stemgent).

iPS cells were transfected with Cas9- and sgRNA-expressing plasmids, and ssODNs by electroporation. Two million cells were resuspended in 100 μl cold BTXpress electroporation buffer (Harvard Apparatus) with 20 μg pCas9_GFP, 5 μg sgRNA plasmid, and 30 μg ssODN (100 bp ssODN, PAGE-purified, IDT). Cells were electroporated at 65 mV for 20 ms in a 1 mm cuvette (Harvard Apparatus). After electroporation, cells were transferred to Geltrex-coated cell culture plates and grown in MEF-conditioned HUESM containing ROCK inhibitor for 2 days. In all transfections, 7889SA-derived iPS cells wild-type at genome-edited loci were used.

HEK293T cells (Life Technologies) were maintained in DMEM with 10% FBS, 2 mM Glutamax and 100 U per ml penicillin and 0.1 mg per ml streptomycin (all Life Technologies) at 37 °C with 5% $CO_2$. HEK293 cells were seeded on 12-well plates at 250,000 cells per ml. When approximately 70% confluent, HEK293 cells were transfected with 800 ng Cas9 plamid, 400 ng sgRNA plasmid and 1 μg ssODN Cells using X-tremeGENE 9 (Roche).

**Fluorescence-activated cell sorting.** All GFP-positive cells, regardless of expression levels, were collected in the Rockefeller University Flow Cytometry Resource Center using a FACSAria II flow cytometer (BD Biosciences). Then 48 h following transfection, cells were resuspended in PBS with 0.5% BSA fraction V solution, 10 mM HEPES, 100 U per ml penicillin, 0.1 mg per ml streptomycin (all from Life Technologies), 0.5 M EDTA, 20 mM glucose, 10 ng per l DAPI in the presence of ROCK inhibitor for iPS cell sorts. For pooled cell next-generation sequencing analysis, 150,000 to 250,000 cells were collected and immediately frozen in liquid $N_2$ for further study. For single-cell derived iPS cell clonal analysis 30,000 GFP$^+$ cells were immediately plated on a 10 cm plate of MEFs in HUESM and ROCK inhibitor following cell sorting.

**Next-generation sequencing analysis of HDR-mediated mutation incorporation.** Genomic DNA was extracted from sorted cells and the genomic region around the CRISPR/Cas9 target site for *APP* and *PSEN1* genes was amplified by PCR with primers positioned outside of the HDR repair template sequence to avoid template amplification for 25 cycles using Q5 polymerase (NEB) according to the manufacturer's protocol (PCR primers listed in Supplementary Table 1). Primers contained sample-specific barcodes. 25 cycles were previously determined to be optimal for exponential amplification of the template as well as visibility for gel extraction (data not shown). To eliminate PCR byproducts and genomic DNA, PCR products were gel purified. 25–100 ng of pooled barcoded PCR products were submitted to the Rockefeller University Genomics Resource Center for targeted MiSeq (Illumina) 300 bp paired-end next-generation sequencing with library preparation using the v3 reagent kit (Illumina).

Data analysis was performed using Galaxy[34,35] (http://usegalaxy.org) or Unix-based software tools listed below (summarized in Extended Data Fig. 8). First, quality of paired-end sequencing reads (R1 and R2 fastq files) was assessed using FastQC[36]. Raw paired-end reads were combined using paired end read merger (PEAR)[37] to generate single merged high-quality full-length reads. Reads with sample-specific forward and reverse barcodes were de-multiplexed using the

FASTX-Toolkit[38] barcode splitter. The barcodes were then trimmed using seqtk (https://github.com/lh3/seqtk). Reads were then filtered by quality (using Filter FASTQ[39]) removing reads with a mean PHRED quality score under 30 and a minimum per base score under 24. Only reads shorter than or equal to the length of the PCR amplicons plus 40 bp (to account for insertions) were considered for analysis.

For the accurate HDR and indel analysis in Fig. 1, reads were filtered to assess the presence of HDR or NHEJ-induced indels. To isolate sequences with HDR, reads were first filtered to remove unedited wild-type reads. Next, HDR reads containing *APP* or *PSEN1* mutations were isolated by matching a 6-nt HDR motif around the pathogenic mutation. HDR reads were then analysed for incorporation of CRISPR/Cas-blocking mutations by matching 6-nt to 8-nt HDR motifs around each mutation and categorized into unique groups of reads containing all possible combinations (32) of CRISPR/Cas-blocking mutations to account for measurable HDR after re-editing (Extended Data Fig. 8b). Each group of reads was then aligned to a corresponding reference sequence using bwa mem[40] (which has been successfully used for this purpose by others[4,41,42]) with option -M to determine the rate of accurate HDR and indel or substitution mutations (Extended Data Fig. 8c). Reads with multiple blocking mutations were analysed separately. Accurate HDR reads were calculated in each group as the percentage of HDR reads without indels. To determine indel frequency, size and distribution, all edited reads from each experimental replicate were combined and aligned, as described above. Indels were then marked at each base using bam-readcount (https://github.com/genome/bam-readcount), quantified in R[43] and plotted using GraphPad Prism.

In all other experiments (all figures except Fig. 1a–e), reads were first filtered for experiment-specific barcode and quality as described earlier (Extended Data Fig. 8a). Next, reads were considered to have HDR if they matched the repair ssODN template plus an additional 3-nt genomic sequence on each side to ensure proper genomic context during HDR and contained the pathogenic mutation and/or CRISPR/Cas-blocking silent mutation (Extended Data Fig. 8d). For all next-generation sequencing experiments, HDR rates were calculated and listed in Extended Data Table 1. *n* values represent independent biological replicates.

To exclude a significant contribution of oligonucleotide synthesis and sequencing errors to our analysis, we sequenced *PSEN1* PCR amplicons from *APP*-edited iPS cells, and *APP*/*PSEN1* repair ssODNs annealed to a complementary ssODN. Errors introduced by sequencing were 2.7% ± 0.1% per 100 bp, and 2.3% ± 1.7% of the 100 bp ssODN sequences contained errors.

**Calculation of optimal distance ranges for homozygous or heterozygous genotypes.** Mutation scan data for *APP* and *PSEN1* loci determined by next-generation sequencing for iPS cells from Fig. 2c were combined to determine single allelic mutation incorporation probabilities $p_a$ as a function of cut-to-mutation distance ($p_a^{mut}$). The probability of wild-type incorporation ($p_a^{wt}$) was determined as ($p_a^{wt} = 1 - p_a^{mut}$). Assuming gene editing and HDR at each allele in a single cell are independent events, we calculated the zygosity probabilities ($p_z$) for each allele combination given two alleles per cell. Specifically, probability of a homozygous, wild-type, and heterozygous zygosity was calculated as $p_z^{mut/mut} = p_a^{mut} \times p_a^{mut}$, $p_z^{wt/wt} = p_a^{wt} \times p_a^{wt}$ and $p_z^{wt/mut} = 2 \times (p_a^{wt} \times p_a^{mut})$, respectively. These calculations were made using the entire range of data derived from Fig. 2c, extrapolated for distance values above 36 and plotted in Fig. 3b as fit curve ± s.d. of raw values.

**RFLP analysis and Sanger sequencing for genotyping of single-cell clones.** To facilitate single-cell clone genotyping, the ssODN HDR templates used for gene editing were designed to introduce a restriction endonuclease motif with the blocking or pathogenic mutation. Genome edited single-cell-derived iPS cell clones grown on MEF-containing 10-cm plates (in HUESM + ROCK inhibitor) were manually picked into a single well of a U-bottom 96-well tissue culture plate in 100 μl HUESM + ROCK inhibitor. Cells were pelleted by centrifugation, and plates were immediately frozen in liquid $N_2$ and stored at −80 °C. Genomic DNA was extracted as previously described[44]. Briefly, cells were resuspended in 25 μl lysis buffer (0.75 μl 10 mg ml$^{-1}$ proteinase K (Ambion), 2.5 μl 10× PCR buffer (Sigma-Aldrich), transferred to 96-well PCR plates and incubated at 55 °C for 4 h. Proteinase K was inactivated by incubating plates at 96 °C for 10 min.

To identify clones with HDR events, the genomic region surrounding the APP$^{Swe}$ or PSEN1$^{M146V}$ loci were amplified by Taq polymerase (Roche) and digested with restriction enzymes to screen for a novel restriction site introduced by the blocking or pathogenic mutation (primers, repair ssODNs and restriction enzymes used are listed in Supplementary Table 1 and 2). Digested DNA was analysed by agarose gel electrophoresis. The zygosity of the pathogenic mutation in clones that had undergone incorporation of the silent CRISPR/Cas-blocking mutations was determined by Sanger sequencing (Genewiz). Bi-allelic HDR rates for single-cell clones were calculated and listed in Extended Data Table 1.

To determine the frequency and distribution of indels in mono-allelic HDR single-cell clones with NHEJ at the other allele, Sanger sequencing reads were separated into single reads for HDR and indel-containing alleles using PolyPeak Parser[45]. Indel-containing reads were then combined into a single FASTA file and analysed for indel distribution by aligning to the reference sequence as described earlier.

**CORRECT.** Re-guide and re-Cas use a two-step gene editing workflow: two million iPS cells were electroporated with sgRNA and Cas9 plasmids. In addition, during the first step, a ssODN containing the intended mutation (M) and a CRISPR/Cas blocking mutation (B) was introduced (MB template). Cas9–eGFP expressing cells were FACS sorted and single-cell iPS cell clones were derived. The presence of B and M mutations was detected by RFLP. A single clone containing homozygous B and M mutations was then expanded for use in the second step of CORRECT. These 'MB iPS cells' were then electroporated with re-sgRNA and wild-type Cas9 plasmids (for re-guide) or wild-type sgRNA and mutant VRER Cas9 plasmids (for re-Cas). In addition, at this step the CORRECT template was provided to remove blocking mutation B. The efficacy of CRISPR/Cas blocking mutation removal was determined by next-generation sequencing. Alternatively, after the second CORRECT step, cells can be plated to derive single-cell scarless 'M iPS cell' clones.

**Off-target analysis.** Gene edited homozygous and heterozygous APP$^{Swe}$ and PSEN1$^{M146V}$ iPS cell lines were tested for off-target editing events predicted for each sgRNA by the Zhang laboratory CRISPR design tool (http://crispr.mit.edu) and the COSMID[46] tool (http://crispr.bme.gatech.edu), which also considers insertions or deletions in the guide RNA target sequence. The top five non-overlapping predicted off-target sites for each sgRNA from each tool were used. The region surrounding each off-target site was PCR-amplified, Sanger sequenced (Genewiz) and compared to the unedited cell line.

**Cortical neuron differentiation.** iPS-cell-derived cortical neurons were generated as previously described[47] with modifications. Specifically, to generate neural precursor cells (NP cells), iPS cells were plated on 12-well tissue culture plates coated with Geltrex (Life Technologies) in MEF-conditioned HUESM with ROCK inhibitor. When cells were 100% confluent, medium was replaced with neural induction (NI) medium (day *in vitro* 0 (DIV0)) and maintained for 8 days. On DIV8 cells were dissociated using Accutase (Life Technologies) and resuspended in NI medium with ROCK inhibitor at 30 million cells per ml. Cells were plated on dried poly-L-ornithine (Sigma-Aldrich) and laminin-coated (Life Technologies) 6-well plates in 10-μl spots. Cells were left to adhere for ∼45 min and NI medium with ROCK inhibitor was added. On DIV10 NI was replaced with neural maintenance (NM) medium. Upon the appearance of neural rosettes, 20 ng ml$^{-1}$ FGF2 was added for 2 days. When neurons started to emerge from rosettes, those were isolated manually after treatment with STEMdiff Neural Rosette Selection Reagent (STEMCELL Technologies) for 1 h. Rosettes were washed and plated on poly-L-ornithine/laminin-coated 6-well plates. Between DIV30 and DIV 36 NPCs were frozen in NM supplemented with 10% DMSO and 20 ng ml$^{-1}$ FGF2.

For cortical neuron maturation, ∼200,000–500,000 NPCs were plated on 24-well poly-L-ornithine/laminin-coated plates and maintained in Neurobasal medium supplemented with B-27 serum-free supplement, 2 mM Glutamax and 100 U per ml penicillin and 0.1 mg per ml streptomycin (all Life Technologies). During the first 7 days after plating, cells were treated with 10 μM DAPT (Sigma-Aldrich) to augment neuronal maturation.

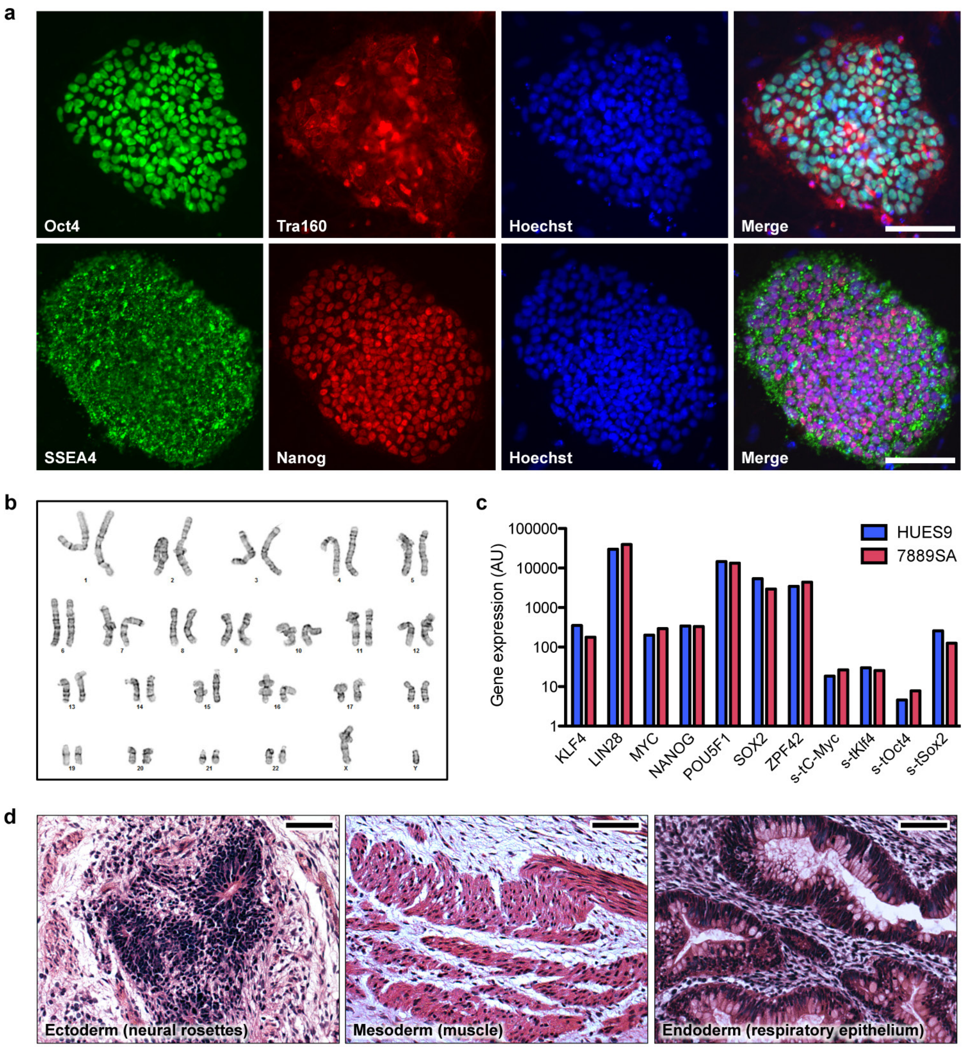
**Cortical neuron characterization.** Canonical neural precursor cell markers (Nestin, Pax6, FoxG1, Otx2) and mature cortical neuronal markers (Tbr1, CTIP2, Satb2) were analysed by immunofluorescence staining at DIV10 and DIV65, respectively. Electrophysiological properties of iPS-cell-derived cortical neurons were assessed between DIV71 and 85 using a submerged recording chamber mounted on an Olympus BX51 microscope equipped for infrared-DIC microscopy. Neurons were perfused with 95% $0_2$/5% $CO_2$ equilibrated ACSF (in mM): 119 NaCl, 2.5 KCl, 1.3 $MgSO_4$, 2.5 $CaCl_2$, 1 $NaH_2PO_4$, 26 $NaHCO_3$ and 11 glucose. Whole-cell patch clamp pipettes (5 MΩ) were filled with (in mM): 123 K-gluconate, 10 HEPES, 0.2 EGTA, 8 NaCl, 2 $Na_2ATP$, 0.3 $Na_3GTP$. Action potentials were elicited by step current injections and recorded in current-clamp mode (−65 mV). Properties (threshold, overshoot) of the largest action potential elicited in each cell were measured. Spontaneous synaptic activity was recorded in voltage-clamp mode (−70 mV). Data was digitized at 10 kHz and recorded using a Multiclamp 700B amplifier and Clampex 10.3.0.2 software (Molecular Devices).

**Amyloid-β measurements.** Aβ was measured in cell supernatant conditioned for 2 days (iPS cells), 3 days (DIV34 neural precursors), or 4 days (DIV72 cortical neurons). Experiments were performed in 3 biological replicates. Supernatants from experiments collected at different time points were frozen at −80 °C. Secreted $A\beta_{1-38}$, $A\beta_{1-40}$ and $A\beta_{1-42}$ were measured with MSD Human (6E10) Aβ V-PLEX kits (Meso Scale Discovery) according to the manufacturer's directions. iPS cell and neuronal total Aβ levels were normalized to total protein levels from cell lysate determined by BCA assay (Pierce).

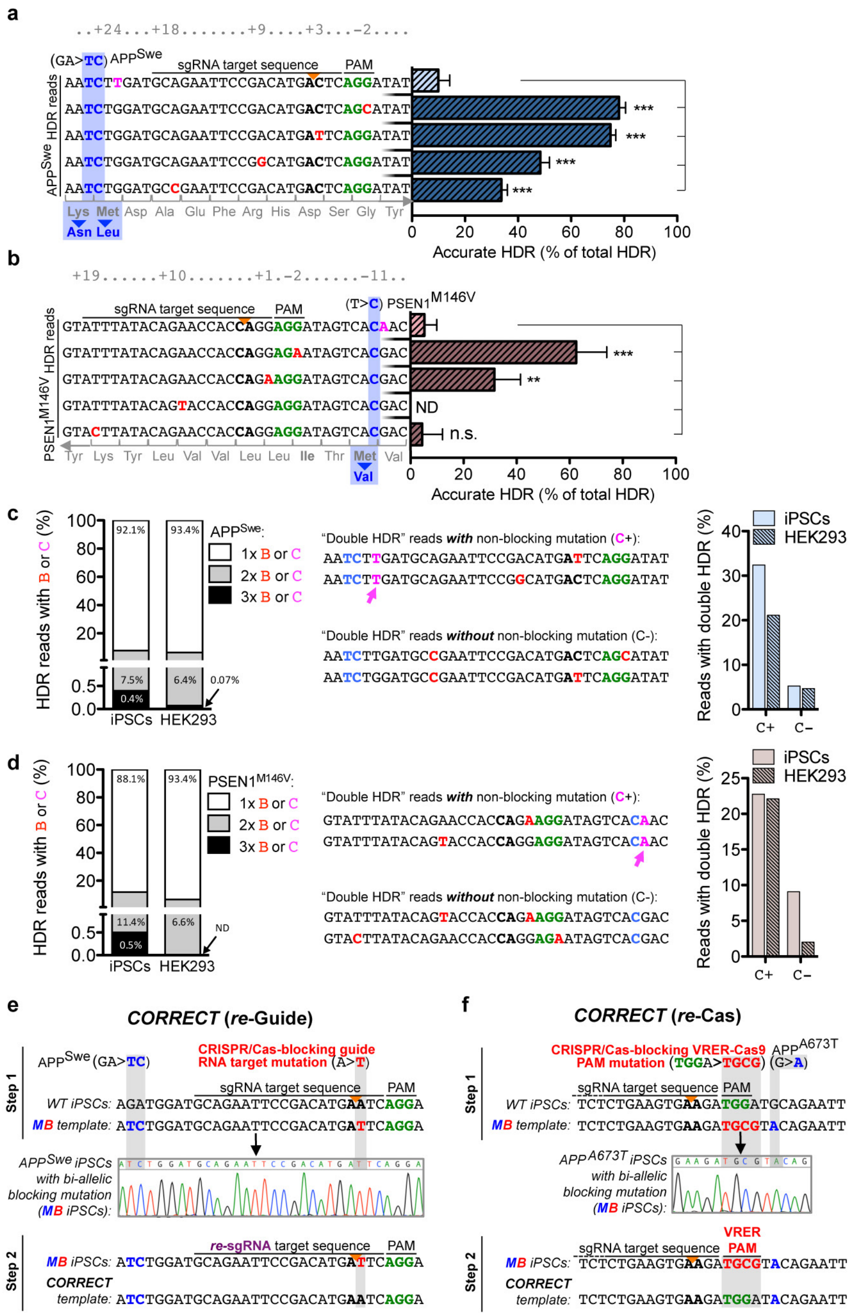
**Surveyor assays.** Genomic DNA was extracted from gene-edited iPS cells as described above. 300–500 bp around the gene-edited locus were amplified by PCR using Herculase II (Agilent) and column purified. PCR amplicons were rehybridized and treated with Surveyor nuclease according to the manufacturer's directions (IDT). Digested DNA was separated on a 4–20% TBE polyacrylamide gel (BioRad) and imaged using SYBR Gold (Life Technologies). Densitometry was performed using Fiji. Per cent indel quantification was based on relative band intensities using the formula $100 \times (1-(1-(b+c)/(a+b+c))^{1/2}$, where $a$ is the undigested PCR product intensity and $b$ and $c$ are the intensities of each cleavage product[48].

**Statistical analysis.** No statistical methods were used to predetermine sample size and the experiments were not randomized. Experimental data was analysed for significance using GraphPad Prism 6. $P < 0.05$ was considered statistically significant. All experiments except the oligonucleotide scan were analysed by one-way ANOVA followed by post-testing with either Tukey's test, if multiple values were compared to each other, or Dunnett's method, if alterations were compared to controls. Similarity of variance was confirmed with Bartlett's test where appropriate. For the oligonucleotide scan, nonlinear regression analysis was performed to fit exponential decay equation model curves to experimental values; $R$ square values were determined to test goodness of fit. To analyse if distance-incorporation relationships were significantly different for genomic loci, the rate constant $k$ was determined for each individual data set and the $k$ values of the two loci were compared using the unpaired $t$-test. The analysis approaches have been justified as appropriate by previous biological studies, and all data met the criteria of the tests. The investigators were not blinded to allocation during experiments and outcome assessment.

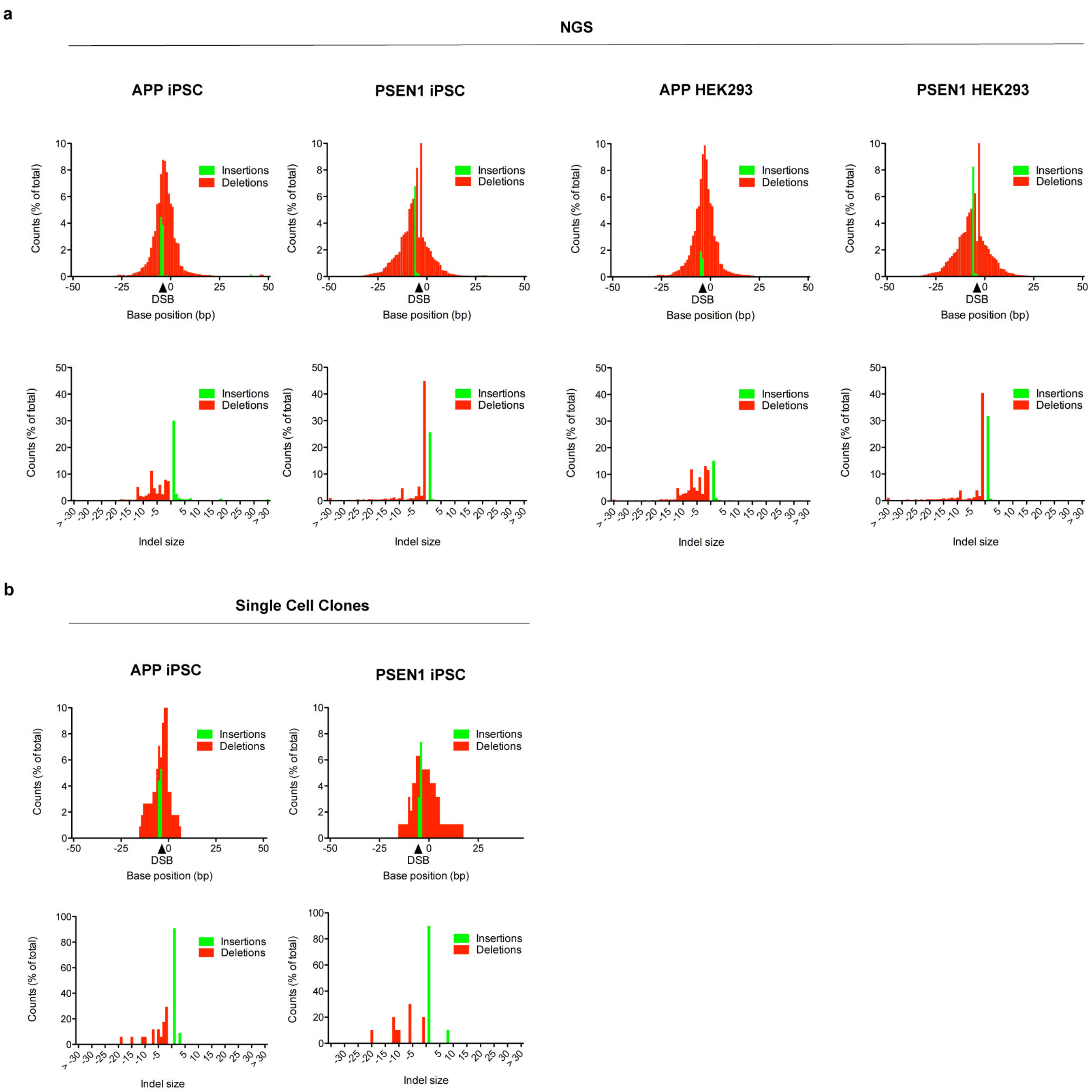

31. Fu, Y. *et al.* High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. *Nature Biotechnol.* **31,** 822–826 (2013).
32. Kahler, D. J. *et al.* Improved methods for reprogramming human dermal fibroblasts using fluorescence activated cell sorting. *PLoS ONE* **8,** e59867 (2013).
33. Cowan, C. A. *et al.* Derivation of embryonic stem-cell lines from human blastocysts. *N. Engl. J. Med.* **350,** 1353–1356 (2004).
34. Goecks, J., Nekrutenko, A. & Taylor, J. & The Galaxy Team. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.* **11,** R86 (2010).
35. Blankenberg, D. *et al.* Galaxy: a web-based genome analysis tool for experimentalists. *Curr. Protoc. Mol. Biol.* **Chapter 19**, Unit 19.10.1–21 (2010).
36. Andrews, S. FastQC: a quality control tool for high throughput sequence data. (http://www.bioinformatics.babraham.ac.uk/projects/fastqc).
37. Zhang, J., Kobert, K., Flouri, T. & Stamatakis, A. PEAR: a fast and accurate Illumina paired-end read merger. *Bioinformatics* **30,** 614–620 (2014).
38. Pearson, W. R., Wood, T., Zhang, Z. & Miller, W. Comparison of DNA sequences with protein sequences. *Genomics* **46,** 24–36 (1997).
39. Blankenberg, D. *et al.* Manipulation of FASTQ data with Galaxy. *Bioinformatics* **26,** 1783–1785 (2010).
40. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26,** 589–595 (2010).
41. Tsai, S. Q. *et al.* Dimeric CRISPR RNA-guided FokI nucleases for highly specific genome editing. *Nature Biotechnol.* **32,** 569–576 (2014).
42. Koike-Yusa, H., Li, Y., Tan, E.-P., Velasco-Herrera, M. D. C. & Yusa, K. Genome-wide recessive genetic screening in mammalian cells with a lentiviral CRISPR-guide RNA library. *Nature Biotechnol.* **32,** 267–273 (2014).
43. R Core Team. R: A language and environment for statistical computing. (R Foundation for Statistical Computing, 2015).
44. Zhu, Z., González, F. & Huangfu, D. The iCRISPR platform for rapid genome editing in human pluripotent stem cells. *Methods Enzymol.* **546,** 215–250 (2014).
45. Hill, J. T. *et al.* Poly peak parser: method and software for identification of unknown indels using Sanger sequencing of polymerase chain reaction products. *Dev. Dyn.* **243,** 1632–1636 (2014).
46. Cradick, T. J., Qiu, P., Lee, C. M., Fine, E. J. & Bao, G. COSMID: a web-based tool for identifying and validating CRISPR/Cas off-target sites. *Mol. Ther. Nucleic Acids* **3,** e214 (2014).
47. Shi, Y., Kirwan, P. & Livesey, F. J. Directed differentiation of human pluripotent stem cells to cerebral cortex neurons and neural networks. *Nature Protocols* **7,** 1836–1846 (2012).
48. Ran, F. A. *et al.* Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity. *Cell* **154,** 1380–1389 (2013).
49. Jonsson, T. *et al.* A mutation in *APP* protects against Alzheimer's disease and age-related cognitive decline. *Nature* **488,** 96–99 (2012).



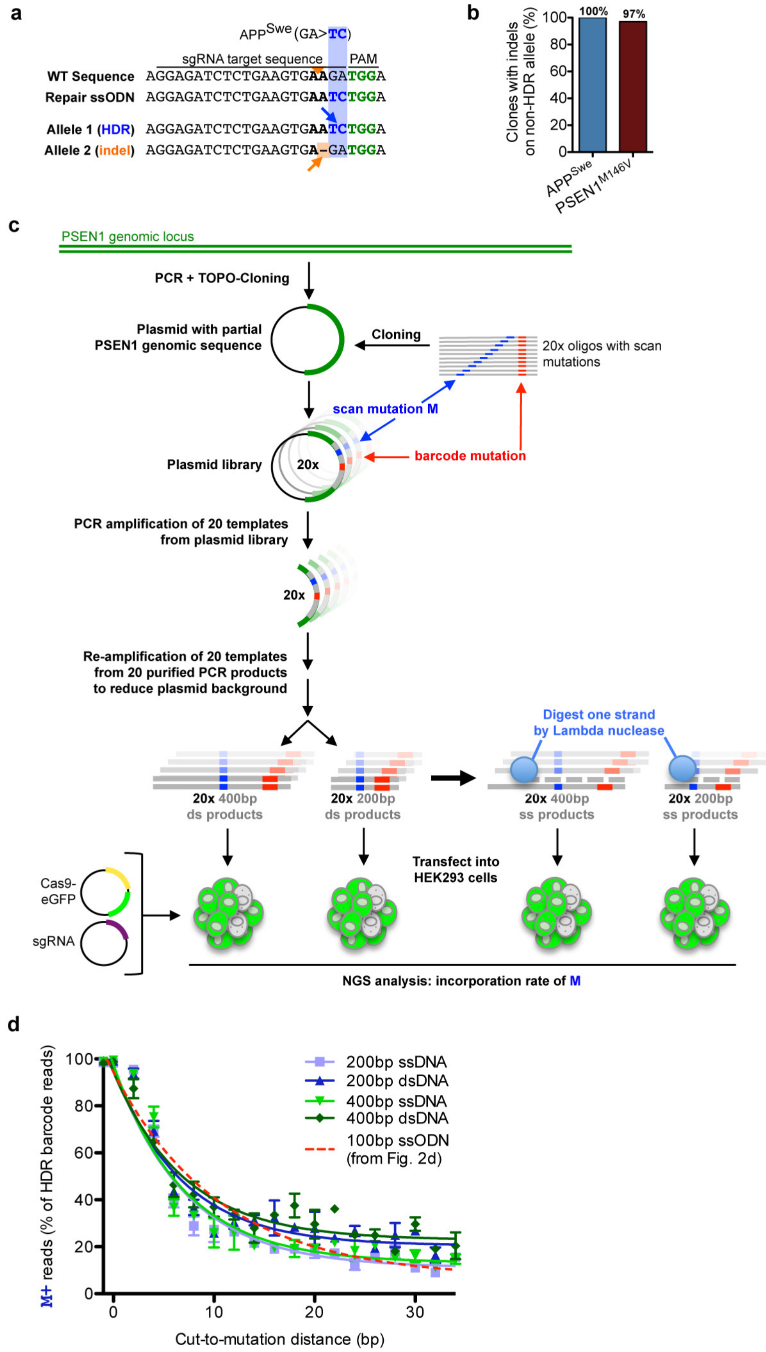

**Extended Data Figure 1 | *In vitro* and *in vivo* characterization of the wild-type 7889SA human iPS cell line. a**, Immunofluorescence staining of pluripotent stem cell markers. **b**, iPS cells possess a normal human male karyotype. **c**, Nanostring expression analysis of pluripotent stem cell genes in reprogrammed iPS cells compared to HUES9. **d**, *In vivo* differentiation and analysis of iPS-cell-derived teratoma containing tissues of all germ cell layers. Scale bars, 100 μm.



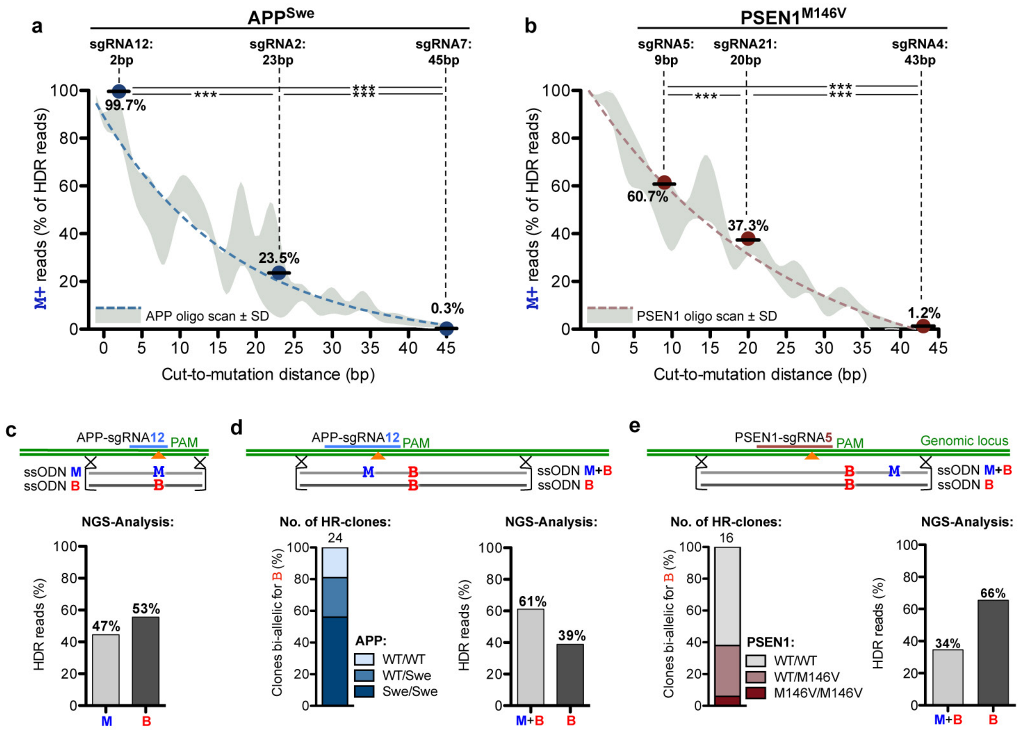

**Extended Data Figure 2 | CRISPR/Cas-blocking mutations increase HDR accuracy by preventing re-editing, are incorporated in multiple rounds of re-editing and can also be applied to scarless editing using CORRECT. a**, **b**, HDR reads from five unpooled templates containing intended pathogenic and CRISPR/Cas-blocking or non-blocking control mutations. Percentages of accurate HDR for reads containing blocking (B) or control (C) mutations at the *APP* (**a**) and *PSEN1* (**b**) locus in HEK293 cells. Values represent mean ± s.e.m. ($n = 3$). ND, not detected. $***P < 0.001$, $**P < 0.01$, one-way ANOVA. **c**, **d**, Proportion of next-generation sequencing reads containing putative single, double, or triple HDR events (left) for *APP* (**c**) and *PSEN1* (**d**). Putative 'double HDR' examples of the most frequent reads that either contain a non-blocking control mutation C with an additional CRISPR/Cas-blocking mutation B, or do not contain C and have two different CRISPR/Cas-blocking mutations (middle). Reads that contain the non-blocking mutation (C+) are more frequently re-edited to incorporate a CRISPR/Cas-blocking mutation ('double HDR') than reads containing a blocking mutation B instead of the non-blocking mutation C (C−). See Fig. 1c for legend. To facilitate data analysis, all replicates were pooled to increase read numbers for rare events. **e**, **f**, Schematics depicting details of the two tested CORRECT approaches: in step 1 of re-guide (**e**), the APP$^{Swe}$ mutation was introduced together with a CRISPR/Cas-blocking guide RNA target mutation, which was then removed again in step 2 using a re-sgRNA specific for the mutated sequence and wild-type Cas9. In step 1 of re-Cas (**f**), the APP$^{A673T}$ mutation was introduced together with a CRISPR/Cas-blocking PAM-altering NGCG mutation, which was then removed in step 2 using the VRER Cas9 variant, which specifically detects the NGCG PAM. We chose to use the very active APP-sgRNA12 to test CORRECT by re-Cas, which was also used in Fig. 3c and 3d to generate APP$^{Swe}$ mutant lines. However, as the APP$^{Swe}$ mutation is located in the target sequence of this sgRNA, it may block re-editing by CRISPR/Cas and could therefore complicate the interpretation of results. We therefore decided to knock-in the protective APP$^{A673T}$ mutation[49] instead, which lies outside of the target sequence. In both cases, the blocking mutations were removed using a CORRECT ssODN repair template, which restored the original sequence at the site of the blocking mutation (which blocks further re-cutting in this step), but retained the intended *APP* mutation. Note that due to repeated editing, CORRECT may increase the probability of off-target effects, but presumably not the number of potential off-target sites, as the same (for re-Cas) or a very similar (for re-guide) guide RNAs are used in both editing steps.



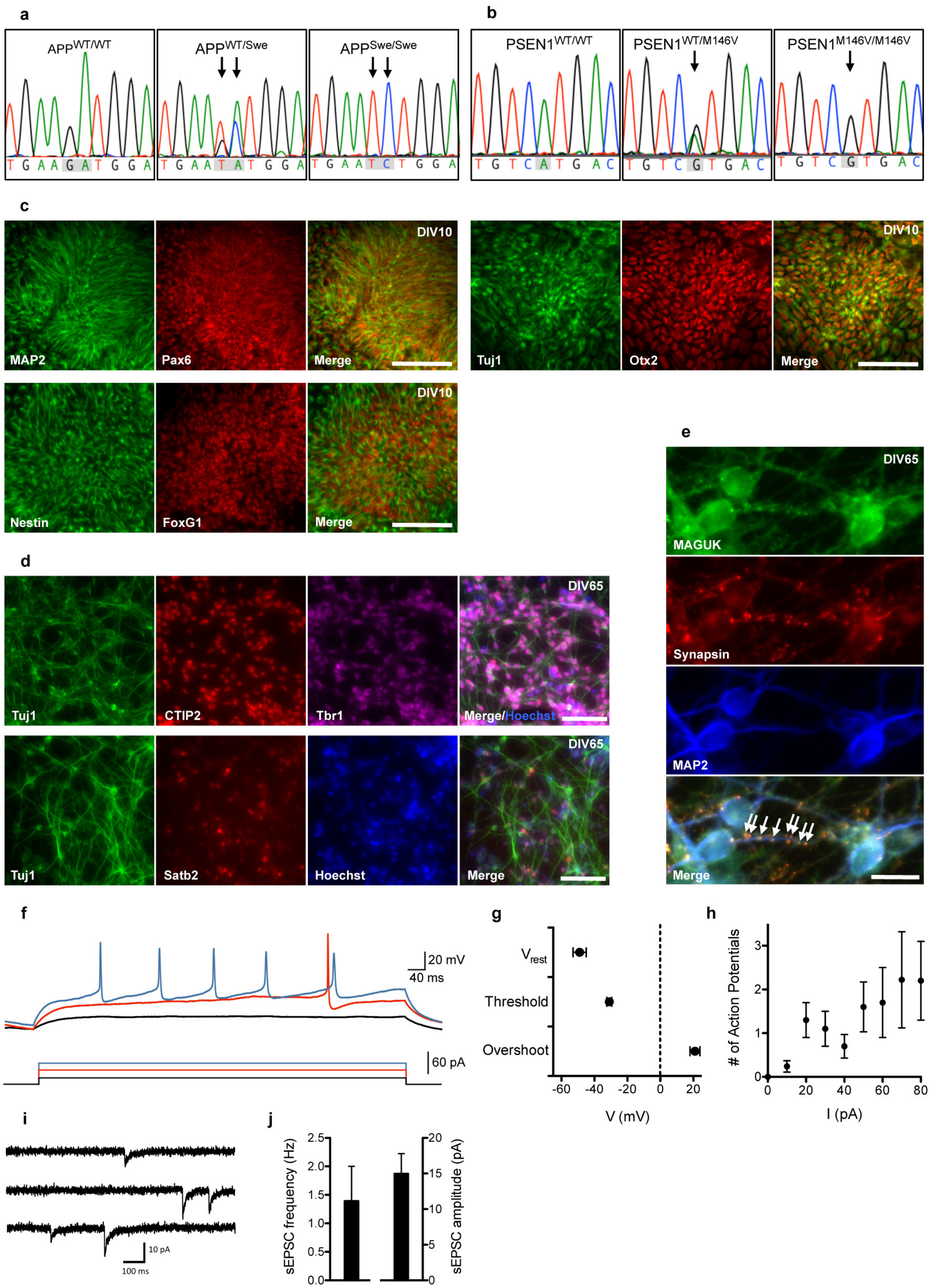

**Extended Data Figure 3 | Analysis of CRISPR/Cas9-induced indels in gene edited iPS cells and HEK293 cells. a**, Plot depicting frequency of indels at each position around the targeted locus in all next-generation sequencing reads with editing events from the analysis shown in Fig. 1. Insertions are plotted at the location where they begin, and deletions are plotted across all deleted base positions (top). Histogram illustrating distribution of indel sizes (bottom). **b**, Indel position (top) and size (bottom) of indel-containing alleles from single-cell clones analysed in Extended Data Fig. 4a, b.



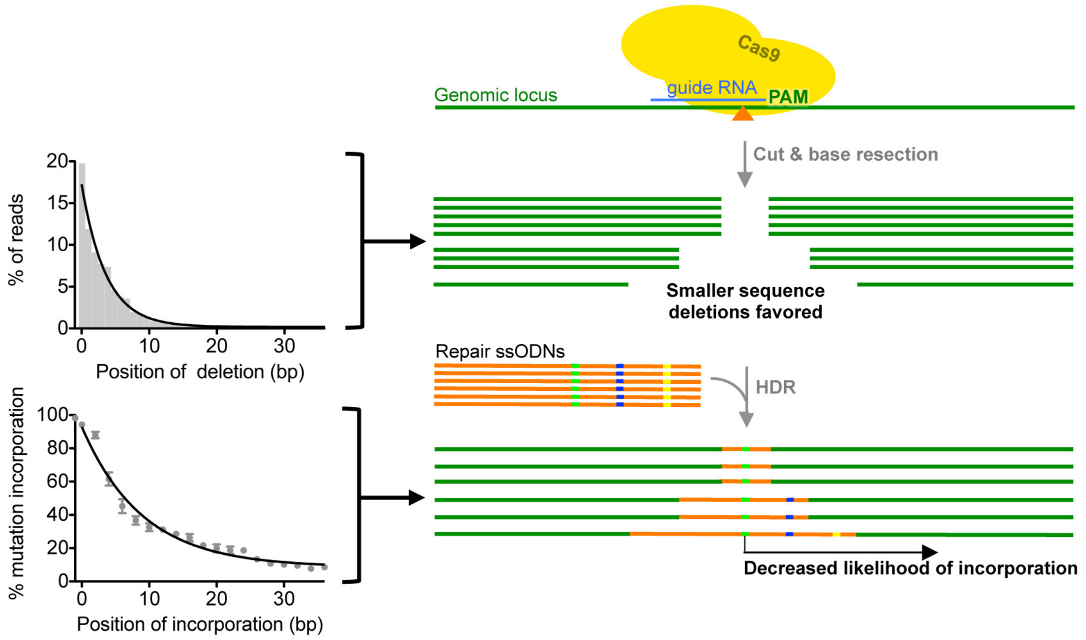

**Extended Data Figure 4 | Heterozygous clones with HDR on one allele almost always contain indels on the non-HDR allele, and longer ssDNA or dsDNA HDR repair templates do not influence mutation incorporation probabilities related to cut-to-mutation distance.** **a**, Sanger sequencing reads of both *APP* alleles of a single-cell clone with mono-allelic HDR (blue arrow). The non-HDR allele is altered by NHEJ in the guide RNA target sequence (orange arrow). **b**, Single-cell clones with HDR on one allele are mostly altered by NHEJ on the non-HDR allele (*APP*, $n = 26$; *PSEN1*, $n = 34$). **c**, Schematic describing the generation of large ssDNA and dsDNA HDR repair templates for the *PSEN1* locus (see Methods for details). **d**, The monotonic relationship between incorporation of intended mutations (M) by HDR and cut-to-mutation distance is not altered by providing longer ssDNA and dsDNA templates ($n = 2$). Red dashed trend line shows previously determined 100-nt oligonucleotide scan result (from Fig. 2d) for comparison.



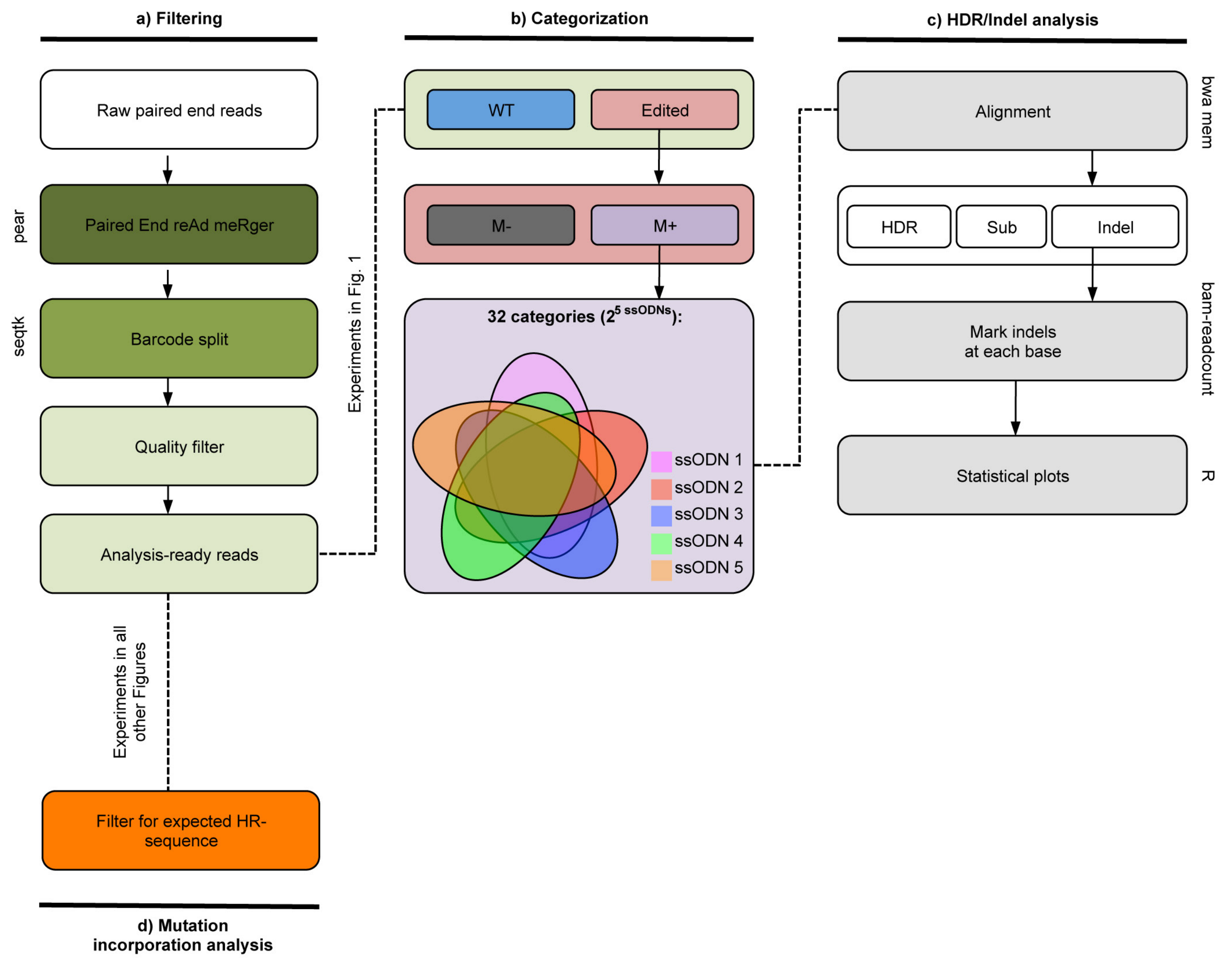

**Extended Data Figure 5 | Mutation incorporation rates at various cut-to-mutation distances follow the distance effect, and mixed repair templates as a strategy to generate heterozygous iPS cell single-cell clones. a**, **b**, Incorporation rate of *APP* and *PSEN1* pathogenic mutations at increasing distance from the cut site targeted by three distinct sgRNA/ssODN pairs is governed by distance. Incorporation rates (solid dots represent mean ± s.e.m., note s.e.m. is too small to be visible, ($n = 3$)) match almost exactly the curves for each locus previously determined by oligonucleotide scan (dashed trend line ± s.d. of raw data from Fig. 2c, d). $***P < 0.001$, one-way ANOVA. **c**, **d**, Mixed ssODN editing approach at the *APP* locus with blocking mutations in one (**c**) or both (**d**) ssODNs (top); zygosity quantification of single-cell clones (**d**, bottom left) and incorporation rates of CRISPR/Cas-blocking mutation B and pathogenic mutation M determined by next-generation sequencing analysis (**d**, bottom right). Note that for the M/B approach in **c**, both oligonucleotides are incorporated at equal levels, as they have similar blocking activities, whereas for the M+B/B approach in **d**, the M+B ssODN is preferentially incorporated, presumably due to a synergistic blocking effect of both M and B. For the clone quantification in Fig. 3d, the rate of wild-type clones was not assessed, because the silent mutation did not introduce a restriction site. However, given the ~50% ssODN incorporation rates determined by deep sequencing, about 25% of HDR clones are predicted to be wild type. **e**, Mixed ssODN editing approach at the PSEN1$^{M146V}$ locus (top). Using an sgRNA with the smallest possible cut-to-mutation distance (PSEN1-sgRNA5), two ssODNs were provided, each containing the same silent PAM-altering CRISPR/Cas-blocking mutation B, but only one containing the pathogenic mutation M. Frequencies of pathogenic mutation genotypes in single-cell clones with bi-allelic HDR of B (bottom left) and incorporation rates of CRISPR/Cas-blocking and pathogenic mutations by next-generation sequencing (bottom right). Note that due to the 9 bp distance to the cleavage site, the incorporation of M is lower than 50% (as expected from the distance effect).




**Extended Data Figure 6 | Characterization of iPS-cell-derived cortical neurons. a**, **b**, Sanger sequencing reads of APP$^{Swe}$ and PSEN1$^{M146V}$ gene edited iPS cell lines. **c**–**e**, Immunofluorescence staining of markers for neural precursors at DIV10 (**c**), cortical neurons at DIV65 (**d**) and functional synapses at DIV65 (**e**). Scale bars; 100 µm (**c**, **d**), 10 µm (**e**). **f**, Evoked action potentials recorded in a neuron current-clamped to −65 mV. **g**, Mean (±s.e.m.) resting membrane potential ($V_{rest}$), action potential threshold and action potential overshoot (DIV 71–85; $n = 18$). Properties of the largest action potential elicited in each cell were measured. **h**, Mean number of evoked action potentials increases with increasing stimulus strength. **i**, Spontaneous synaptic activity recorded in a neuron voltage-clamped to −70 mV. **j**, Mean (± s.e.m.) frequency and amplitude of spontaneous excitatory postsynaptic currents (sEPSCs) (DIV 71–85; $n = 8$).




**Extended Data Figure 7 | Possible mechanism underlying the distance effect for HDR-mediated mutation incorporation with CRISPR/Cas9.** CRISPR/Cas9 causes a DSB at a genomic locus, which leads to variable size deletions or strand resections in different cells. Genomes with small deletions or resections are more common than large ones, which is reflected in the distribution of deleted bases after NHEJ (top left). During HDR, only the part of the repair template overlapping this deletion may be used, which results in fewer mutations incorporations more distal to the cleavage site (bottom left, data pooled for *APP* and *PSEN1* from Fig. 2d).




**Extended Data Figure 8 | Next-generation sequencing data analysis pipeline for HDR and indel detection. a**, For all next-generation sequencing experiments, raw forward and reverse paired next-generation sequencing reads were first merged to obtain single high-quality reads (tool: PEAR), de-multiplexed to separate experiment-specific barcoded reads (seqtk) then filtered to remove low-quality reads. **b**, For experiments using pooled oligonucleotides containing CRISPR/Cas-blocking mutations (displayed in Fig. 1), reads were separated into wild-type (WT) and edited reads, which were then filtered to include only reads that had incorporated the pathogenic mutation (M+) (that is, containing a pathogenic and CRISPR/Cas-blocking mutation). To account for multiple HDR events after re-editing, reads were then separated into 32 unique categories covering every possible combination of CRISPR/Cas-blocking mutations. **c**, Reads were aligned (bwa mem) and accurate HDR (perfect alignment) or indel distribution was reported (bam-readcount, R). For analysis in Extended Data Fig. 2c, d, reads that had incorporated multiple CRISPR/Cas-blocking mutation were separately analysed. **d**, For the mutation incorporation analyses performed in all other figures reads were filtered for the expected sequence and counted.

**Extended Data Table 1 | List of HDR rates determined by next-generation sequencing and single-cell clone analysis**

**a**

| Figure | Locus | sgRNA | Template type | Cell type | % HDR of total reads (SD) | % HDR of edited reads (SD) |
|---|---|---|---|---|---|---|
| 1d | $APP^{Swe}$ | 2 | ssODN | iPSC | 4.3 (0.7) | 10.4 (1.2) |
| 1d | $APP^{Swe}$ | 2 | ssODN | HEK293 | 4.5 (0.2) | 10.6 (0.8) |
| 1e | $PSEN1^{M146V}$ | 22 | ssODN | iPSC | 2.2 (0.2) | 2.8 (0.2) |
| 1e | $PSEN1^{M146V}$ | 22 | ssODN | HEK293 | 1.2 (0.1) | 2.6 (0.4) |
| 1i | $APP^{Swe}$ | 2 | ssODN | iPSC | 3.8 (0.1) | 15.7 (1.7) |
| 1k | $APP^{A673T}$ | 12 | ssODN | iPSC | 0.3 (0.1) | 3.5 (0.4) |
| ED2a | $APP^{Swe}$ | 2 | ssODN | HEK293 | 4.9 (0.1) | 8.2 (0.2) |
| ED2b | $PSEN1^{M146V}$ | 22 | ssODN | HEK293 | 1.2 (0.1) | 1.9 (0.1) |
| 2c | $APP^{Swe}$ | 2 | ssODN | iPSC | 2.1 (0.7) | 5.7 (3.1) |
| 2c | $PSEN1^{M146V}$ | 22 | ssODN | iPSC | 3.2 (1.5) | 4.1 (2.0) |
| 2d | $APP^{Swe}$ | 2 | ssODN | HEK293 | 4.1 (0.2) | 9.6 (0.4) |
| 2d | $PSEN1^{M146V}$ | 22 | ssODN | HEK293 | 4.3 (0.2) | 9.1 (0.1) |
| ED4d | $PSEN1^{M146V}$ | 22 | 200 ssDNA | HEK293 | 6.0 (0.04) | 13.8 (0.1) |
| ED4d | $PSEN1^{M146V}$ | 22 | 200 dsDNA | HEK293 | 3.6 (0.1) | 9.5 (0.3) |
| ED4d | $PSEN1^{M146V}$ | 22 | 400 ssDNA | HEK293 | 5.1 (0.05) | 11.6 (0.3) |
| ED4d | $PSEN1^{M146V}$ | 22 | 400 dsDNA | HEK293 | 3.1 (0.1) | 8.3 (0.1 |
| ED5a | $APP^{Swe}$ | 12 | ssODN | iPSC | 5.9 (0.3) | 10.6 (0.4) |
| ED5a | $APP^{Swe}$ | 2 | ssODN | iPSC | 6.7 (1.5) | 10.9 (0.3) |
| ED5a | $APP^{Swe}$ | 7 | ssODN | iPSC | 0.4 (0.1) | 1.1 (0.2) |
| ED5b | $PSEN1^{M146V}$ | 5 | ssODN | iPSC | 1.8 (0.8) | 4.8 (0.1) |
| ED5b | $PSEN1^{M146V}$ | 22 | ssODN | iPSC | 2.0 (0.3) | 3.4 (0.6) |
| ED5b | $PSEN1^{M146V}$ | 4 | ssODN | iPSC | 1.6 (0.3) | 6.8 (1.4) |
| ED5c | $APP^{Swe}$ | 12 | ssODN | iPSC | 3.5 (ND) | 5.7 (ND) |
| ED5d | $APP^{Swe}$ | 12 | ssODN | iPSC | 3.6 (ND) | 8.6 (ND) |
| ED5e | $PSEN1^{M146V}$ | 5 | ssODN | iPSC | 2.1 (ND) | 4.4 (ND) |

**b**

| Figure | Locus | sgRNA | Template type | Picked clones | Bi-allelic HDR clones | % bi-allelic HDR clones |
|---|---|---|---|---|---|---|
| 3c | $APP^{Swe}$ | 12 | ssODN | 720 | 24 | 3.3 |
| 3c | $APP^{Swe}$ | 2 | ssODN | 912 | 20 | 2.2 |
| 3c | $APP^{Swe}$ | 7 | ssODN | 1623 | 6 | 0.4 |
| 3c | $PSEN1^{M146V}$ | 5 | ssODN | 912 | 22 | 2.4 |
| 3c | $PSEN1^{M146V}$ | 21 | ssODN | 960 | 21 | 2.2 |
| 3c | $PSEN1^{M146V}$ | 4 | ssODN | 1056 | 20 | 1.9 |
| 3d | $APP^{Swe}$ | 12 | ssODN | 768 | 16 | 2.1 |
| ED5d | $APP^{Swe}$ | 12 | ssODN | 1056 | 24 | 2.3 |
| ED5e | $PSEN1^{M146V}$ | 5 | ssODN | 192 | 16 | 8.3 |

**Extended Data Table 2 | Off-target analysis of knock-in APP$^{Swe}$ and PSEN1$^{M146V}$ iPS cell lines**

**a**

| Gene | gRNA | Tool | ID | Sequence | Type | Mis-match | Location | Strand | COSMID score | ZHANG score | Indels in het |
|---|---|---|---|---|---|---|---|---|---|---|---|
| APP | 2 | | | GCAGAATTCCGACATGACTCAGG | Target | 0 | 21:25897598-25897620 | - | | | |
| APP | 2 | COSMID | OT1 | ACA-AATTCCAACATGACTCTGG | Del 17 | 2 | 16:48796036-48796057 | - | 1.5 | | None |
| APP | 2 | COSMID | OT2 | CCA-AATTCCAACATGACTCTGG | Del 17 | 2 | 1:147436251-147436272 | + | 1.5 | | None |
| APP | 2 | COSMID | OT3 | CCA-AATTCCAACATGACTCTGG | Del 17 | 2 | 3:103798317-103798338 | - | 1.5 | | None |
| APP | 2 | COSMID | OT4 | TCA-AATTCCAACATGACTCTGG | Del 17 | 2 | 5:62074737-62074758 | - | 1.5 | | None |
| APP | 2 | COSMID | OT5 | GCA-AATTCCAACATGACTCTGG | Del 17 | 2 | X:4446208-4446229 | + | 2.68 | | None |
| APP | 2 | ZHANG | OT6 | TCACAATTCTGACATGACTCAGG | No indel | 3 | 9:80909384-80909406 | - | | 2.525 | None |
| APP | 2 | ZHANG | OT7 | GCAGGACTCCAACATGACTCAGG | No indel | 3 | 12:63304976-63304998 | + | | 0.964 | None |
| APP | 2 | ZHANG | OT8 | GCAGAATCCAGACATGACCCAAG | No indel | 3 | 3:108345664-108345686 | + | | 0.838 | None |
| APP | 2 | ZHANG | OT9 | TGTGAATTCCAACATGACTCAAG | No indel | 4 | 2:156152304-156152326 | + | | 0.795 | None |
| APP | 2 | ZHANG | OT10 | GGTGAATACCAACATGACTCAAG | No indel | 4 | 1:207331941-207331963 | + | | 0.782 | None |

**b**

| Gene | gRNA | Tool | ID | Sequence | Type | Mis-match | Location | Strand | COSMID score | ZHANG score | Indels in homo |
|---|---|---|---|---|---|---|---|---|---|---|---|
| APP | 12 | | | GGAGATCTCTGAAGTGAAGATGG | Target | 0 | 21:25897623-25897642 | - | | | |
| APP | 12 | COSMID | OT1 | TG-GATCTCTGAAGTGAAGATGG | Del 18 | 1 | 1:25967504-25967525 | + | 0.78 | | None |
| APP | 12 | COSMID | OT2 | AGTGA-CTCTGAAGTGAAGATGG | Del 15 | 2 | 9:21442538-21442559 | + | 0.99 | | None |
| APP | 12 | COSMID | OT3 | GG-GTTGTCTGAAGTGAAGAGGG | Del 18 | 2 | 20:32252244-32252265 | + | 1.08 | | None |
| APP | 12 | COSMID | OT4 | GGACTTC-CTGAAGTGAAGAAGG | Del 13 | 2 | 10:1347959- 13479615 | - | 1.14 | | None |
| APP | 12 | COSMID | OT5 | GAAAATCT-TGAAGTGAAGAAGG | Del 12 | 2 | 6:102732573-102732594 | + | 1.16 | | None |
| APP | 12 | ZHANG | OT6 | GGAAATCTCTAAAGTGAAGAAAG | No indel | 2 | 6:21801477-21801499 | + | | 3.934 | None |
| APP | 12 | ZHANG | OT7 | GGACCTCACTGAAGTGAAGAAGG | No indel | 3 | 1:85276298-85276320 | - | | 2.426 | None |
| APP | 12 | ZHANG | OT8 | TAAGATTTCTGAAGTGAAGAAAG | No indel | 3 | 11:128019181-128019203 | - | | 1.737 | None |
| APP | 12 | ZHANG | OT9 | GAAGATCCCTGGAGTGAAGACAG | No indel | 3 | 4:84241317-84241339 | - | | 1.384 | None |
| APP | 12 | ZHANG | OT10 | GCAGATCTCGGAAGTGAAGCTAG | No indel | 3 | 2:464569-464591 | + | | 1.374 | None |

**c**

| Gene | gRNA | Tool | ID | Sequence | Type | Mis-match | Location | Strand | COSMID score | ZHANG score | Indels in het | Indels in homo |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| PSEN1 | 5 | | | TGTTGTCATGACTATCCTCCTGG | Target | 0 | 14:73173656-73173678 | + | | | | |
| PSEN1 | 5 | COSMID | OT1 | TG-TATCATGGCTATCCTCCCGG | Del 18 | 2 | 17:71454517-71454538 | + | 1.55 | | None | None |
| PSEN1 | 5 | COSMID | OT2 | TGCTGACATGA-TATCCTCCAGG | Del 9 | 2 | 17:22640654-22640675 | + | 1.67 | | None | None |
| PSEN1 | 5 | COSMID | OT3 | TATTGTCAT-AATATCCTCCTGG | Del 11 | 2 | 12:26052306-26052327 | - | 1.94 | | None | None |
| PSEN1 | 5 | COSMID | OT4 | TGTTTTAATGAC-ATCCTCCAGG | Del 8 | 2 | 14:19342489-19342510 | + | 2.03 | | None | None |
| PSEN1 | 5 | COSMID | OT5 | TGTTTTAATGAC-ATCCTCCAGG | Del 8 | 2 | 15:20767818-20767839 | + | 2.03 | | None | None |
| PSEN1 | 5 | ZHANG | OT6 | TGATGTCATCACTATCCTCCAAG | No Indel | 2 | 13:62696807-62696829 | + | | 6.438 | None | None |
| PSEN1 | 5 | ZHANG | OT7 | TTTTTTCCTGACTATCCTCCCAG | No Indel | 3 | 9:36047217-36047239 | + | | 2.543 | None | None |
| PSEN1 | 5 | ZHANG | OT8 | TGATGTCCTGACTATCCTCAAAG | No Indel | 3 | 4:153813094-153813116 | + | | 1.422 | None | None |
| PSEN1 | 5 | ZHANG | OT9 | AGATATCATCACTATCCTCCTAG | No Indel | 4 | 12:105761175-105761197 | + | | 1.299 | None | None |
| PSEN1 | 5 | ZHANG | OT10 | TTATCTAATGACTATCCTCCCAG | No Indel | 4 | 12:29848003-29848025 | - | | 0.905 | None | None |

**a–c**, List of properties of the five most similar off-target sites predicted each for APP-sgRNA2 used for heterozygous APPSwe lines (**a**), APP-sgRNA12 used for homozygous APPSwe lines (**b**) and PSEN1-sgRNA5 used for both heterozygous and homozygous PSEN1M146V lines (**c**) using COSMID or the Zhang laboratory CRISPR design tool. Red bases indicate sequence differences from target sequence. No off-target indels were identified.

**CORRIGENDUM**

doi:10.1038/nature16515

# Corrigendum: Dissecting a circuit for olfactory behaviour in *Caenorhabditis elegans*

Sreekanth H. Chalasani, Nikos Chronis, Makoto Tsunozaki, Jesse M. Gray, Daniel Ramot, Miriam B. Goodman & Cornelia I. Bargmann

*Nature* **450,** 63–70 (2007); doi:10.1038/nature06292
corrigendum *Nature* **451,** 102 (2008); doi:10.1038/nature06540

We have discovered that Figs 1, 2, 3 and 5 and Supplementary Figs 1–5 and 7 of this Article were generated from calcium imaging data that included a number of duplicated or mislabelled movie files. A full reanalysis of source files indicated that 163 of the 851 unique movie files (19%) were invalid. We have regenerated all figures using only valid movies, which changed the *n* values for many data points. 36/41 calcium imaging experiments were not significantly affected, and remained statistically robust after the reanalysis. The properties of $AWC^{ON}$, AIB and AIY neurons were fully supported, as were the effects of odours, mutations and most conditions. However, for 5/41 experiments the corrected *n* values included only two or three movies, so the conclusions should be considered preliminary. These results include a limited analysis of the $AWC^{OFF}$ sensory neurons (Fig. 2c–e) and two AIB neuron time points (Fig. 3c and d). The Supplementary Information to this Corrigendum shows corrected versions of all affected figures and the associated figures in the Supplementary Information of the original Article, using only valid movies, and a detailed list of re-analysed experiments with *n* values. We apologise for these errors.

**Supplementary Information** is available in the online version of the Corrigendum.

CORRIGENDUM

doi:10.1038/nature16538

# Corrigendum: Discovery of Atg5/Atg7-independent alternative macroautophagy

Yuya Nishida, Satoko Arakawa, Kenji Fujitani, Hirofumi Yamaguchi, Takeshi Mizuta, Toku Kanaseki, Masaaki Komatsu, Kinya Otsu, Yoshihide Tsujimoto & Shigeomi Shimizu

*Nature* **461,** 654–658 (2009); doi:10.1038/nature08455

In Supplementary Fig. 19a of this Letter, the 'no treatment' panels for Stx7 contain incorrect data, owing to an error in image placement during figure preparation. The Supplementary Information to this Corrigendum shows the corrected Supplementary Fig. 19a, and the raw data from which we produced the corrected panels. This error does not affect the description, interpretation or conclusions of the Letter.

**Supplementary Information** is available in the online version of the Corrigendum.

# CORRECTIONS & AMENDMENTS

CORRIGENDUM

doi:10.1038/nature16968

## Corrigendum: DDX5 and its associated lncRNA Rmrp modulate TH17 cell effector functions

Wendy Huang, Benjamin Thomas, Ryan A. Flynn, Samuel J. Gavzy, Lin Wu, Sangwon V. Kim, Jason A. Hall, Emily R. Miraldi, Charles P. Ng, Frank Rigo, Sarah Meadows, Nina R. Montoya, Natalia G. Herrera, Ana I. Domingos, Fraydoon Rastinejad, Richard M. Myers, Frances V. Fuller-Pace, Richard Bonneau, Howard Y. Chang, Oreste Acuto & Dan R. Littman

*Nature* **528**, 517–522 (2015); doi:10.1038/nature16193

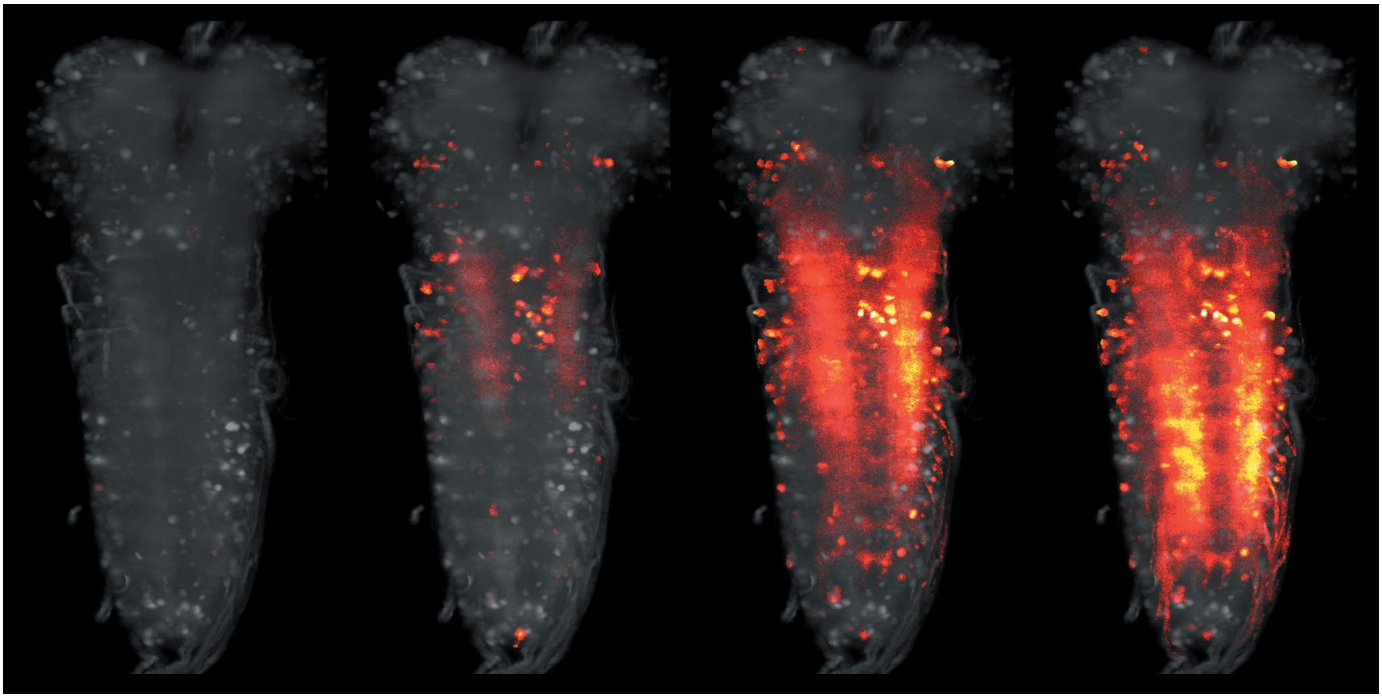
In this Article, author ‘Frank Rigo’ was incorrectly listed with a middle initial; this has been corrected in the online versions of the paper.

TOOLBOX

# THE STRUGGLE WITH IMAGE GLUT

*Experiments that generate millions of images have forced scientists to find new ways to store and share terabytes of experimental data.*

W.C. LEMON *ET AL. NATURE COMMUN.* **6,** 7924 (2015)



**Neurons fire in a fruit-fly larva: a single experiment to track this activity produces millions of images like these.**

**BY JEFFREY M. PERKEL**

As the fruit-fly larva wriggles forwards in the video, a crackle of neural activity shoots up its half-millimetre-long body. When it wriggles backwards, the surge undulates the other way. The 11-second clip, which has been watched more than 100,000 times on YouTube, shows the larva's central nervous system at a resolution that almost captures single neurons. And the experiment that created it produced several million images and terabytes of data.

For developmental biologist Philipp Keller, whose team produced the video at the Howard Hughes Medical Institute's Janelia Research Campus in Ashburn, Virginia, such image-heavy experiments create huge logistical challenges. "We've spent probably about 40% of our time during the past 5 years simply investing in computational methods for data handling," he says. The problem isn't so much storing images — data storage is cheap — but organizing and processing the images so that other scientists can make sense of them and retrieve what they need.

The 'image glut' challenge is becoming an increasing burden for researchers across the biological and physical sciences. Here, Keller and scientists in two other fields — astronomy and structural biology — explain to *Nature* how they are tackling the problem.

## MAPPING THE SUN

Somewhere in geosynchronous orbit above Las Cruces in New Mexico, the Solar Dynamics Observatory (SDO) traces a figure-of-eight in the sky. The satellite keeps a constant watch on the Sun, recording its every hiccup and burp with an array of three instruments that photograph the Sun through ten filters, record its ultraviolet output and track its seismic activity. Those data are then beamed to a ground station below. The SDO produces "something like 1.5 terabytes of image data a day", says Jack Ireland, a solar scientist at ADNET Systems, a NASA contractor in Bethesda, Maryland. According to NASA, this amount of data is equivalent to about 500,000 iTunes songs.

To help researchers to stay on top of those images, the ADNET team at NASA, with the European Space Agency, developed the Helioviewer website (helioviewer.org) for browsing SDO images — rather like Google Maps for the Sun, says Ireland — as well as a downloadable application (jhelioviewer.org).

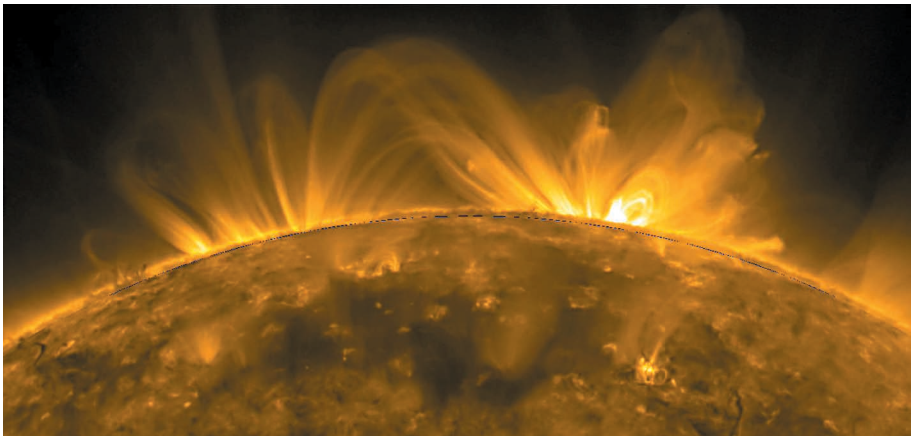
Researchers and astronomy enthusiasts using these tools view not the original data, but instead a lower-resolution representation of them. "We have images of the data," Ireland explains, "not the data itself."

The original SDO scientific images are each 4,096 × 4,096 pixels square and about

12 megabytes (MB) in size. They are taken every 12 seconds, and tens of millions have been collected — a data archive of several petabytes (PB), and growing (1 PB is 1 billion MB, or 1,000 TB). To make images accessible to users, every third image is compressed to 1 MB and made available through Helioviewer.

Users can jump to any particular time since the SDO launched in 2010, select a colour filter and retrieve the data. They can then zoom in, pan around and crop the images, and string them together into movies to visualize solar dynamics. Users create about 1,000 movies a day on average, Ireland says, and since 2011, at least 70,000 have been uploaded to YouTube.

Once they have selected an individual image or cropped area, such as the region around a particular solar flare, users can still download it in its original high resolution. They can also download the complete archive of smaller 1-MB images if they want: but at 60 TB and counting, that process could take weeks.


NASA/SDO

Activity on the Sun seen by NASA's Solar Dynamics Observatory, which gathers 1.5 terabytes of data a day.

## FASTER FILE FORMATS

For Keller's developmental-biology group at the Janelia Research Campus, posting their data online for outsiders to access isn't such a concern. If others request it, the team can share images using specialist file-transfer tools, or simply by shipping hard drives. First, however, the team must manage and sort through images that stream off the lab's microscopes at the rate of a gigabyte each second. "It's a huge challenge," Keller says.

Keller's lab uses microscopes that fire sheets of light into the brains and embryos of small organisms such as fruit flies, zebrafish and mice. These have been genetically modified so that their cells fluoresce in response — allowing the team to image and track each cell in 3D for hours. To store its data, the lab has spent around US$140,000 on file servers that provide about 1 PB of storage.

The highly structured organization of the millions of images on those servers keeps the team sane. Each microscope stores its data in its own directory; files are arrayed in a tree that describes the date a given experiment was done, what model organism was used, its developmental stage, the fluorescently tagged protein used to visualize the cells, and the time that each frame was taken. The lab's custom data-processing pipeline was constructed to act on that organization, Keller says.

Yet the directories don't contain the JPEG image files with which most microscopists are familiar. The JPEG format compresses image file sizes, making them easier to process and transfer, but it is relatively slow at reading and writing those data to disk, and is inefficient for 3D data. Keller's microscopes collect images so fast that he needed a file format that could compress images as efficiently as JPEG, but that could be written and read much faster. And because the lab often works on isolated subsets of the data, Keller needed a simple way to extract specific spatial locations or time points.

Enter the Keller Lab Block (KLB) file format, developed by Keller and his team. This chops up image data into chunks ('blocks'), which are compressed in parallel by multiple computer processors[1]. That triples the speed at which files can be read and written, so KLB can compress file sizes just as well as the JPEG format, if not better.

In theory, Keller says, KLB files could be used on commercial digital cameras or on any system that requires rapid data access. KLB source code is freely available, and the lab has made tools and file converters for the MATLAB programming environment and for an open-source image-analysis package called ImageJ, as well as for some commercial packages. Researchers using commercial microscopes could employ the format too, says Keller; he calls it "straightforward" to convert data to KLB files for long-term storage and use.

## SHARING RAW DATA

Biologists who take pictures to determine molecular structures also generate vast amounts of image data. And one technique that is growing in popularity — and hence, generating more data — is cryoelectron microscopy (cryoEM).

CryoEM users fire electron beams at a flash-frozen solution of proteins, collect thousands of images and combine these to reconstruct a 3D model of a protein with near-atomic resolution. Most of these reconstructions are less than 10 GB in size, and researchers deposit them in the Electron Microscopy Data Bank (EMDB) — but not the raw data used to create them, which are some two orders of magnitude larger than the resulting models. The EMDB simply was not set up to handle them, says Ardan Patwardhan, who leads the EMDB project for the Protein Data Bank in Europe (PDBe) at the European Bioinformatics Institute (EBI) near Cambridge, UK. As a result, reproducibility suffers, Patwardhan says: without access to raw data, researchers can neither validate others' experiments nor develop new analysis tools.

In October 2014, the PDBe launched a pilot solution: a database of raw cryoEM data called the Electron Microscopy Pilot Image Archive (EMPIAR), also led by Patwardhan. Only data sets for structures deposited in the EMDB are allowed, he says; otherwise, users might be tempted to use the database as a data dump.

EMPIAR currently contains 49 entries averaging 700 GB apiece. The largest is more than 12 TB, and the total collection weighs in at about 34 TB. "We have space available to grow into the petabyte range," Patwardhan says. Users download about 15 TB of data per month in total.

Downloading such large amounts of data presents its own problems: the standard protocol used to transfer files between computers, called FTP, struggles with large data sets; connection loss is common, and download times can slow significantly over long distances. Instead, the EBI has paid for EMPIAR users to access two high-speed file-transfer services, Aspera and Globus Online, both of which transfer data at the rates of "a few terabytes per 24 hours", Patwardhan says. The EBI — which also uses these services to transfer large genomics data sets — pays for its side of the transaction. The cost to the EBI of providing Aspera can be many tens of thousands of dollars per year, he says.

The EMPIAR raw data has already proved its worth. Edward Egelman, a structural biologist at the University of Virginia in Charlottesville, co-authored a study[2] of the structure of an aggregated, filament-like protein called MAVS — which was at odds with another, earlier model of the protein[3]. Egelman proved the earlier structure was incorrect by downloading and reprocessing the raw data set[4]. EMPIAR's grant runs out in 2017, but Patwardhan says that cryoEM researchers have told him they already consider EMPIAR a necessity, and want 'pilot' taken out of the archive's name. "They feel that this should be considered a vital archive for the community — which is nice to hear," he says. ■

1. Amat, F. *et al. Nature Protoc.* **10,** 1679–1696 (2015).
2. Wu, B. *et al. Mol. Cell.* **55,** 511–523 (2014).
3. Xu, H. *et al. eLife* **3,** e01489 (2014).
4. Egelman, E. H. *eLife* **3,** e04969 (2014).

**US ACADEMIA** Higher salaries, fewer tenured faculty **p.135**

**SCIENTIFIC SOCIETIES** Women are under-represented **p.135**

**NATUREJOBS** For the latest career listings and advice **www.naturejobs.com**

LOC/SCIENCE FACTION/GETTY

Marie Curie (second from right) and her daughter Irène (second from left) both followed careers in science.

RELATIONSHIPS

# Scions of science

*Relatives in the same career bring advantages — and challenges — for junior researchers.*

**BY AMBER DANCE**

For Irina Kareva, maths is a family affair. The theoretical biologist and her parents, both mathematicians, find maths conferences convenient gathering points. At a 2011 meeting in Vancouver, Canada, Kareva and her father presented a study that the three of them had co-authored, and then went on a hike together.

"Working with my parents is awesome," says Kareva, who lives in Boston, Massachusetts, and is seeking a job. "It brings us together."

Children generally forge a different path from their parents', but science certainly runs in some clans, as it did in the Curie family, whose members laid claim to five Nobel prizes. A 2013 survey commissioned by the family-history website Ancestry.co.uk found that 7% of respondents end up in the same career as a parent. Ismail Onur Filiz and Lada Adamic, social scientists at Facebook, found in an unpublished study of career inheritance that 3.2% of users who listed a career in science had at least one parent who did, too. (The latter number may be an underestimate, they add, because they could analyse only users whose parents were also on Facebook and listed their jobs.)

"I suspect there are many science families," says Celia Schiffer, director of the Institute for Drug Resistance at the University of Massachusetts Medical School in Worcester. Schiffer, who has several scientists in her family, estimates that half of her colleagues have parents who are scientists or engineers. The succession rates for science careers may be particularly high in some cultures, adds Jinsong Liu, a biochemist at the Chinese Academy of Sciences' Guangzhou Institutes of Biomedicine and Health. China, for example, sees a relatively high proportion of offspring choose a parent's career.

Being the junior member of a family of scientists has clear perks: budding researchers have a great way to find out the workings of the academic enterprise, and can benefit from introductions and connections from their parents (see 'Family ties'). But there are downsides, too: young scientists generally want to distinguish themselves from their relatives and avoid any suggestion that their career success is the result of unfair advantage. They also may face extra pressure to succeed if they feel they need to live up to a parent's reputation, or if others expect them to achieve similar levels of success.

And ultimately, it is important for children or relatives of a researcher — especially an ▶

eminent one — to remember that their life is their own and that they, not their parents, are responsible for their career progress. "A big thing I had to overcome, for myself, is that I'm not just an extension of them," says Kareva.

#### THRILLS AND SPILLS

Growing up with a scientist, one sees both the positives, including the thrill of discovery, and the negatives, such as pressure to get funding. Kevin Gardner, director of the Structural Biology Initiative at the City University of New York's Advanced Science Research Center, says that his teenaged daughters can tell when he's stressing about a grant. But they also get to see that his delight in his work outweighs the headaches, and he's already helped to connect his daughter with an expert to advise on her plant-biology projects.

Students can often parlay their family-scientist connection into a low-level lab job. Although some institutions don't allow faculty members to hire their children, positions can sometimes be found in a neighbouring lab. "I call them 'parking-lot jobs,'" says Carolyn Jensen, director of the Eberly College of Science Academic Advising Center at the Pennsylvania State University (Penn State) in University Park. Her father — who was a biochemist at Penn State when she was growing up — sometimes cornered a colleague on the way to the car park to say, "Oh, my daughter's looking for a summer job." These sorts of positions can be a crucial first step for a scientist-in-training, says Sue Biggins, a cell biologist at the Fred Hutchinson Cancer Research Center in Seattle, Washington. They provide valuable exposure to the process of science and a chance to develop key skills.

Once they move on from lab chores to real research, young scientists ought to find independent mentors. It would be hard to be an effective adviser when the relationship includes a family link, Biggins says, and she points out that for future jobs, researchers will need honest, unbiased letters of recommendation.

Kareva, for one, was careful to assert her independence during her PhD at Arizona State University in Tempe. Although her parents acted as unofficial co-advisers for her thesis, she says that it was important and a source of pride for her to include some dissertation chapters that they had nothing to do with.

Now that Kareva is looking for a job, she's ready to embrace her parents' professional network. Given the competitive nature of science, researchers agree that there's no shame in using a relative's network. "It is critical, in this day and age, for people to seek out opportunities wherever they can," advises Belinda Huang, former executive director of the National Postdoctoral Association in Washington DC.

Scientists who share a workplace with a parent or relative should be prepared for potentially awkward moments. Biostatistician Paul Edlefsen and his mother, public-health researcher Nicole Urban, both work at the Fred Hutchinson centre. Occasionally, he will encounter someone at work who knew him as a child, and who comments on how proud his mother must be of him. "The best thing to do is to take the compliment as intended," he says. But Edlefsen tries to draw a line between family and career — he avoids seeking help on his work from his mother, he says, as a way to maintain that divide.

Family members who work at the same institution must take care to avoid conflicts of interest. Bill Berquist, a gastroenterologist and clinical researcher at Lucile Packard Children's Hospital Stanford in Palo Alto, California, kept out of the selection of fellows in his department when his daughter applied. Now, they work together at the hospital. And Maryanne Large, a physicist at the University of Sydney in Australia, earned her undergraduate degree at the University of Sydney in Australia, where her father was a faculty member. The physics department arranged her father's teaching schedule so that he would never grade her.

Even if relatives are based at different institutions but work in the same field, family connections can create tangles. Grant reviews must be independent: committee members at the US National Institutes of Health (NIH) are required to leave the review panel if a close family member applies. The situation differs in the case of permanent reviewers, who serve for years. If there is a potential conflict, the NIH dictates that the grant application must be reviewed by a different department. That means that the proposal might be evaluated differently from others in the field, Biggins says.

#### FOR THE PARENTS

Having a parent in the same career is an advantage — but only if the parents provide the right level of encouragement. Successful scions of science say that their parents never pushed them into their career — rather, mum or dad just made the job look appealing.

Tagging along on work trips makes a big impression, says David Sabatini, a biochemist at the Whitehead Institute in Cambridge, Massachusetts, who travelled with his cell-biologist father. "He seemed to have an interesting life, with interesting and smart people in it," Sabatini says. Sabatini has already taken his own five-year-old son along to a presentation in the Turks and Caicos Islands. (The boy liked the talk, but admonished his father for not explaining things well enough.)

Parents can help their children to get started on a scientific career, but should make sure that science is the right course for them. If the child's school does not offer much career counselling, the parents can point their children to online assessments such Career Driver Online by SkillScan to figure out whether science matches their skills and interests.

A positive family influence is especially important for girls and women in areas of science in which they are under-represented. Elizabeth Larson, an undergraduate student who is majoring in physics and English at the University of Virginia in Charlottesville, has only a few other women in her physics classes — just as her geologist mother was surrounded by mostly male peers during her education. Larson says that she's faced casual sexism — for example, a professor who noted that only his male students had played with electromagnets as kids — and talking with her mum has helped her to work through it.

As scientists climb the career ladder, savvy parents can offer advice on selecting mentors or a place to work, as well as on job prospects in different fields, and even technical advice on specific projects. Jennifer Leeds, head of antibacterial discovery at the Novartis Institute for Biomedical Research in Emeryville, California, says that as her older son was looking around different universities, she called in every

### FAMILY TIES

*How to make relationships work to your advantage*

If your family already boasts a successful scientist or three, it's possible to take advantage of their knowledge or connections while still forging an independent career. Here are some key tips.

- Don't feel bad about tapping into your relative's network. Everyone has one — yours just happens to share some DNA.
- If you're working with a relative, make sure to lay and reinforce ground rules. For example, say, "Please talk to me as you would a collaborator," recommends theoretical biologist Irina Kareva in Boston, Massachusetts.
- Keep it professional at work. (You might need to refer to mum as 'Dr Smith' around patients.)
- Be careful to avoid even the appearance of a conflict of interest; step back from any committees or decisions that involve your relative.
- Be sure to seek diverse mentors outside your family.
- Don't stress about living up to the career of a hotshot parent or relative.
- Realize that some people might think that your successes are down to your relative, rather than your own merit. If they bring it up, explain that even if your parent helped you to make a contact, the successes after that were your own. **A.D.**

COURTESY KAREVA FAMILY

Theoretical biologist Irena Kareva (centre) is inspired by her mathematician parents.

connection she had so that he could meet faculty members and learn more about the institutes. Jensen recalls that her father made subtle suggestions about which professors to take courses from when she was an undergraduate at Penn State; for those he didn't recommend, he'd say, "He's not particularly interested in undergraduate education." Later on, he gave her the name of the person she applied to for a teaching position at the university, although she earned the position on her own merit. And when Gardner started working in the field of optogenetics, he recalls that his father, a retired laser development engineer, would help him to decide what laser systems to use and how to couple the lasers to fibres, which was valuable as he got started.

> ***"What matters is what you do and how you prove yourself once you're given that opportunity."***

Parents should keep a light touch, advises Berquist, so that their offspring can feel independent and think creatively on their own. Three of his four children worked in his office as teens, but he let others supervise them. Now that they are grown — two work at the Children's Hospital and one is pursuing a medical degree — he is careful not to hover. "I try to avoid being very critical, unless I'm asked," he says.

**HIGH EXPECTATIONS**

It can be a bit tough to follow the career of a parent, particularly if she or he is a superstar. Stuart Cahalan originally found it disheartening to think that he might never measure up to his father, biophysicist Michael Cahalan, a department chair at the University of California, Irvine, and a member of the US National Academy of Sciences. "Every kid wants to be better than their parent, or at least as good," says Stuart, a postdoc in biology at the Scripps Research Institute in La Jolla, California.

When the younger Cahalan was being interviewed for graduate school, faculty members quickly recognized his father's name, he says, and he realized that in science, he would always be Michael Cahalan's son. In time, he came to accept that fact. And his knowledge now exceeds that of his father's in some areas. Recently, he proofread and commented on one of his father's grant applications.

Although scientists may stay independent of their parents, others may not necessarily believe that to be the case. David Sabatini's brother Bernardo, a neurobiologist at Harvard Medical School in Boston, Massachusetts, says that he's heard occasional quips about the easy path he must have had because their father was well known by those higher up at the university because he chaired the cell-biology department at New York University. "My response is always the same," says Bernardo. "Even if somebody has help getting in the door, what matters is what you do and how you prove yourself once you're given that opportunity."

But the occasional bit of snark or envy is no reason not to follow in a parent's footsteps, be it the field the child chooses or the institution where he or she lands, scientists say.

Nepotism can't get one ahead too much in science anyway, they say, because researchers are evaluated independently in grant and paper reviews. Success comes from passion and hard work, not pedigree. "It's all about following your heart," says Schiffer. ■

**Amber Dance** *is a freelance writer in Los Angeles, California.*

**FACULTY POSITIONS**

## Tenure figures tumble

The number of US faculty members who have tenure or are on the tenure track is falling, according to a report by the American Association of University Professors in Washington DC. Over the past 40 years, the proportion of the academic labour force that is in a full-time tenured position has shrunk by one-quarter, and the proportion in tenure-track posts has halved, reports *Higher Education at a Crossroads*. In 2014, the study found, 21% of faculty appointments were full-time tenured and 41% were part-time. On average, male professors earned more than female professors in full-time positions at every rank and across all types of institution. Overall, positions in New England paid the most, whereas those in Iowa, Kansas, Minnesota, Missouri, Nebraska, North Dakota and South Dakota paid the least. The report also found that part-time appointees were less likely to conduct long-term research and experiment with teaching methods and course content. Citing a correlation between lower student-graduation rates and increases in the number of part-time and non-tenure-track positions, the association calls for institutions to convert part-time, non-tenure positions into tenure-track posts.

**ACADEMIES**

## Diversity drive

Women represent an average of 12% of the memberships of academic science societies worldwide, finds a report from the InterAcademy Partnership (IAP): The Global Network of Science Academies. *Women for Science: Inclusion and Participation in Academies of Science* examined the membership of 69 national science societies around the world, and found that women comprise 14–16% of academies whose members are concentrated in the biological, medical and social sciences. In maths and engineering societies, women average 5–6% of membership. Female representation on academy governing boards, however, average 20%. Just 40% of the societies said that they have a gender policy or strategy to increase female participation in academy activities. The report recommends that IAP member academies collect and report data annually on membership and activities. It also suggests that academies create committees to establish strategies that will boost gender equality in membership and governance.

# THE MUSEUM OF NOTHING

*But is it art?*

BY ANNA ZUMBRO

The first thing they ask at the Museum of Nothing is that you remove your contact lenses. They even have cases and little bottles of solution, free of charge, although considering how much tickets cost, I suppose it's a stretch to say they're giving you anything for free.

"I have astigmatism," I say when the guard asks if I'm wearing smart lenses. "They're for a prescription." They're also for augmentation, of course, like everybody else's, which is the real reason the museum requires you to leave them behind.

The guard is unmoved. "It's the Museum of *Nothing*," she says pointedly. "It's not like you'll be missing out."

She meets my eyes, allowing my lenses to get a good read on her. Her name is Wanda Richardson and she lives on Webster Street. She dropped out of high school at sixteen, joined a gang and spent nine months in prison for conspiracy to commit robbery, although all she did was act as lookout while her friends bashed in the door to the electronics shop. A church mission helped her turn her life around, until she had an affair with the mission's pastor. She found a second chance — or third, perhaps — as a guard at the museum.

Wanda's eyes narrow. She knows I'm looking at her bio feed that's being projected by my lenses, that I'm processing her past. She isn't wearing lenses, though. Apparently even the guards have to follow the rules here. Seems foolish. You'd think the guards would know better whom to watch if they had lenses feeding them information on the patrons. But then, if it really is the Museum of Nothing, I guess there's nothing to steal.

I write my name on the bag Wanda gives me and remove my lenses at the small sink next to her desk. Her face blurs. She points at the entrance to the exhibition, and I put my hand out in case my eyes fail me.

The gallery has white walls and wooden floors, just like the other galleries I've been to, the ones with Da Vincis and Kahlos and Wyeths and all the rest. But here, there are no paintings, no sculptures, nothing but four white walls and a plaque next to the door. I squint to read the plaque: *Please do not touch the artwork*. And I start to laugh. I can't help it. There's nothing here.

ILLUSTRATION BY JACEY

"So, what brought you to the museum?" asks a man about my height. He smiles as he says this. I think he's amused *with* me, not at me. But it's hard to tell. His features are blurry, almost abstract. I shiver as I realize that without my lenses I don't know his name or his background, don't know if he volunteers at the soup kitchen every Thursday or is on the run from a triple-murder charge. I can't remember the last time I spoke to someone without knowing their name or background.

"I guess I had to see if it was real," I say. "A museum of nothing, you know, I thought it was a joke."

"It did make you laugh." The man's voice sounds friendly. Is it? Can I trust myself to know?

"Why did you come?" I ask.

He gestures at the empty wall in front of him. "I came to see the art."

Now I know something's not right. I step backward. "Sure," I say, humouring him. "It's like nothing I've ever seen."

"No, no. Don't misunderstand," he rushes. "I'm not saying the wall is art. But what is art for, if not to make you think? All day long we're fed information. We never have to figure anything out. Don't you see? This place makes you think. It sparks the imagination. That's why it's art."

I follow his gaze to the wall. If there's any texture in the paint, I can't make it out. All I can think is that I wish I had my contacts back. The world feels strange without them.

"I never thought of myself as old-fashioned before," I say, "but I've always preferred art that I can actually see."

The man steps closer to the wall and leans forward. "Perhaps you should look again," he says. "I recognize this one. *Spilled Milk on White Marble*."

"Is it, now?" I say. "I had confused it for that other masterpiece, *Invisible Ghosts Dancing in Fog*."

We both laugh. His is a nervous laugh, which sets me at ease. I wonder what he suspects about me.

"I'm Tara," I say.

He holds out his hand, then grasps mine when I reach and miss. "Duke," he says. "It's a pleasure to meet you."

I'm not Tara and, somehow, I know he's not Duke. My head hurts from the eyestrain. Still, I'd like to stay for a while. 'Duke' and I will have to leave separately if we don't want to shatter the illusion.

But right now, it doesn't matter. In here, we can paint ourselves a new identity, Picasso morphing from blue to rose to crystal. We can be anyone we want when no one knows otherwise. Far from a Museum of Nothing, in this place, we're both the artists and the art. ■

**Anna Zumbro** *lives in Washington DC. Her stories have appeared in* Cricket, Daily Science Fiction, Grievous Angel *and other publications.*

NATURE.COM
Follow Futures:
@NatureFutures
go.nature.com/mtoodm

nature OUTLOOK

5 May 2016
Supplement to Nature
Publishing Group journals

# RESEARCH COMMERCIALIZATION


## Transforming discoveries into technological innovations

natureOUTLOOK

RESEARCH COMMERCIALIZATION

5 May 2016 / Vol 533 / Issue No 7601


Cover art: John Harwood

**Editorial**
Herb Brody, Michelle Grayson, Richard Hodson, Jenny Rooke

**Art & Design**
Wesley Fernandes, Mohamed Ashour, Andrea Duffy

**Production**
Karl Smart, Ian Pope, Matthew Carey

**Advertising**
Janet Cen, Helen Hill, Kris Kent, Neil MacMillan, Stella Yan, Nicole Yu

**Marketing**
Nicole Jackson

**Project Manager**
Anastasia Panoutsou

**Art Director**
Kelly Buckheit Krause

**Publisher**
Richard Hughes

**Editorial director, partnership media**
Stephen Pincock

**Chief Magazine Editor**
Rosie Mestel

**Editor-in-Chief**
Philip Campbell

A thirst for a fuller and more nuanced understanding of the Universe is a powerful motivation for research. But pursuit of commercial success is also a compelling driver. The ability of these forces to interact and reinforce one another is propelling scientific enterprise forward.

Universities, industry and government, each with their own objectives, cultures and strengths, are locked together in a synergistic embrace that is fuelling a push to extract commercial value from academic research (see page S6). Companies are under pressure to uncover the next business-sustaining product before their competition, and universities are being pushed to deliver a pay-off for their research outlay. As a result, academic institutions are improving their ability to transfer science into the commercial sector (S13). Many are nurturing entrepreneurs, and, in turn, benefiting from the spin-off companies that they launch (S10).

Despite the speed and ease of communication offered by the Internet, researchers still congregate in geographical clusters, suggesting that there is an advantage to proximity that modern technology cannot yet overcome (S40).

Governments are motivated by economic growth, reflected by their status as the principle funders of science and technology. But the economic value of research, and what it means to get a return on research and development, is a matter of discussion (S20). And with science funded by the super-rich on the rise, governments may find themselves facing competition for primacy in funding (S43).

Different parts of the world face different challenges. Despite their research prowess, China (S32) and Australia (S22) have struggled with commercial translation, whereas Europe still needs to better align its policies, habits and business cultures with the goal of efficiently capitalizing on the fruits of cutting-edge research (S30 and S47).

**Herb Brody**
*Supplements Editor*

*Nature Outlooks* are sponsored supplements that aim to stimulate interest and debate around a subject of interest to the sponsor, while satisfying the editorial values of *Nature* and our readers' expectations. The boundaries of sponsor involvement are clearly delineated in the Nature Outlook Editorial guidelines available at go.nature.com/e4dwzw

**CITING THE OUTLOOK**
Cite as a supplement to *Nature*, for example, *Nature* **Vol. XXX**, No. XXXX Suppl., Sxx–Sxx (2016).

**VISIT THE OUTLOOK ONLINE**
This *Nature Outlook* supplement can be found at http://www.nature.com/nature/outlook/research-commercialization
It features all newly commissioned content as well as a selection of relevant previously published material.

All featured articles will be freely available for 6 months.

**SUBSCRIPTIONS AND CUSTOMER SERVICES**
Site licences (www.nature.com/libraries/site_licences): Americas, institutions@natureny.com; Asia-Pacific, http://nature.asia/jp-contact; Australia/New Zealand, nature@macmillan.com.au; Europe/ROW, institutions@nature.com; India, npgindia@nature.com. Personal subscriptions: UK/Europe/ROW, subscriptions@nature.com; USA/Canada/Latin America, subscriptions@us.nature.com; Japan, http://nature.asia/jp-contact; China, http://nature.asia/china-subscribe; Korea, www.natureasia.com/ko-kr/subscribe

**CUSTOMER SERVICES**
Feedback@nature.com

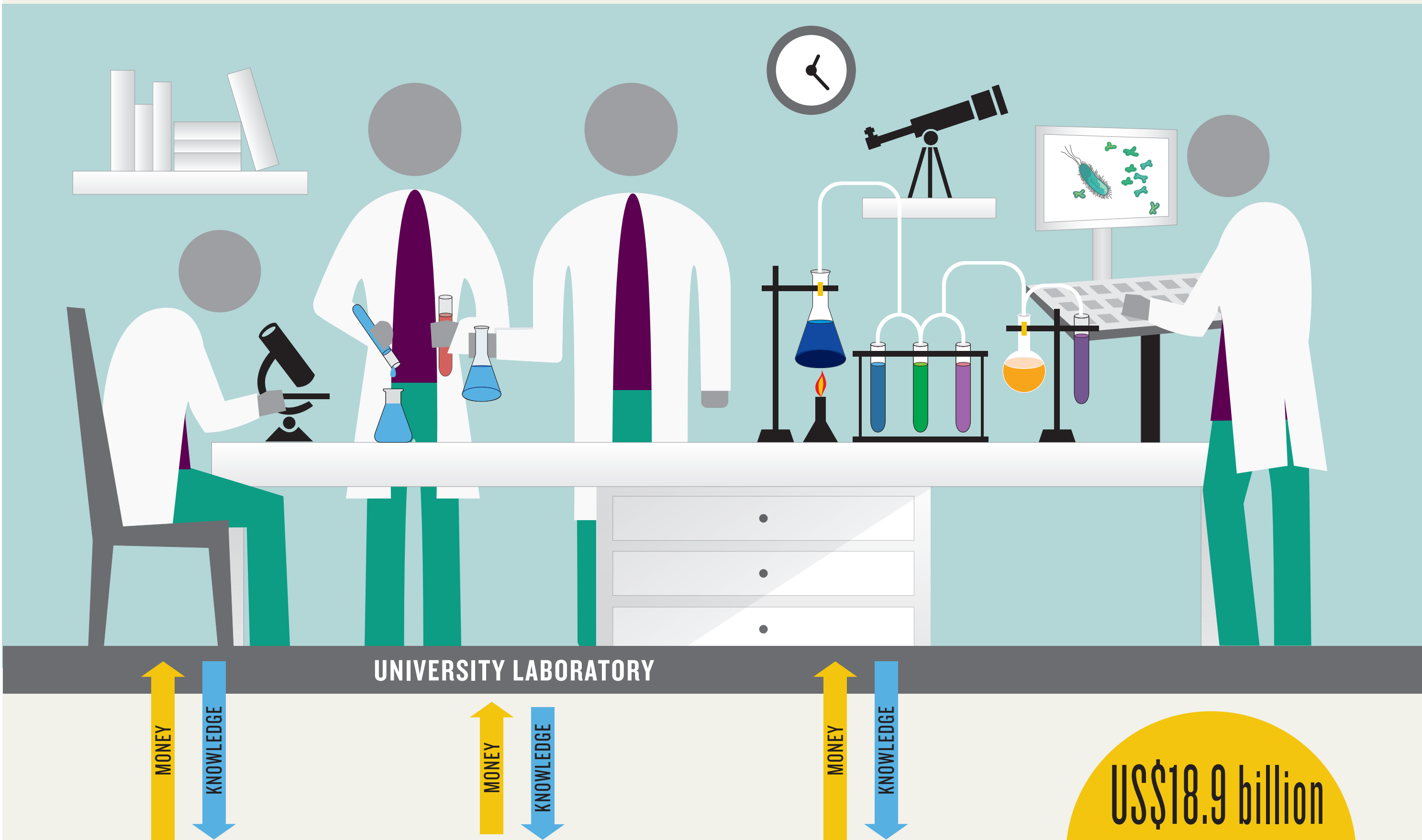

## CONTENTS

# CASHING IN ON SCIENCE

*University research powers innovation and economic development. Countries with intensive research and development (R&D) programmes differ in their approach to turning lab studies into commercial enterprises. By* **Alla Katsnelson**, *infographic by* **Mohamed Ashour**.

## CENTRAL COG

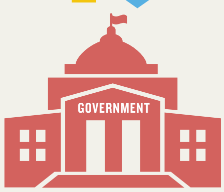
University research drives the innovation ecosystem by generating inventions, patents and licensing agreements, and by spurring the creation of spin-off companies. The funding for this research comes from multiple sources. Academic researchers also generate income or contribute to knowledge commercialization through contract and collaborative research with companies, as well as through consulting.





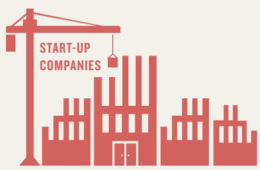
### INNOVATION GENERATION

Money from industry supports a significant amount of university research through direct funding, research collaboration and contract work. This research can generate patents and licensing revenue for the university.




### DEEP INVESTMENT

Government is a key pillar of research commercialization. In the United States, government funds, generated partly by taxes, support some 60% of the research conducted by universities. The dividends from this societal investment are not just the result of commercialization of university-generated inventions, but also knowledge generation more broadly.




### NEW BUSINESS

Scientific or technological achievements with strong commercial potential can spur the creation of start-up companies. Firms can form a variety of agreements with the university at which the research was conducted. The company, for example, may pay licensing fees, patent fees or royalties on the sales of the product.

**US$18.9 billion**

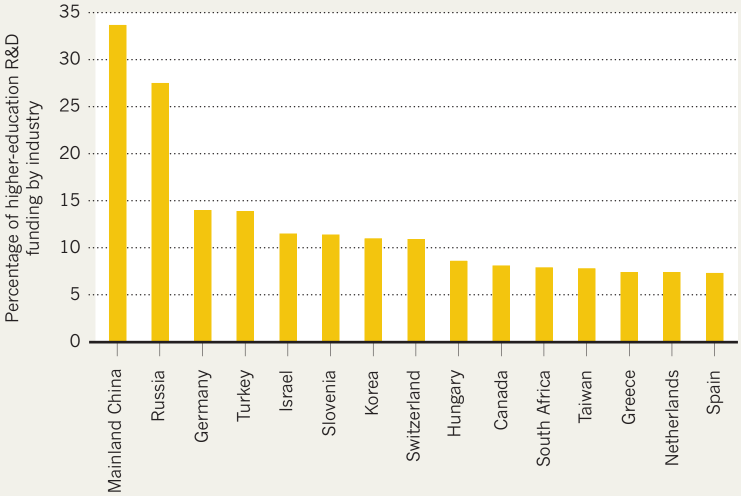
Total amount that industry has provided for higher-education R&D worldwide (using purchasing-power parities at current prices)[1].

**15,623**

Total number of international patent applications worldwide from public research universities[1].

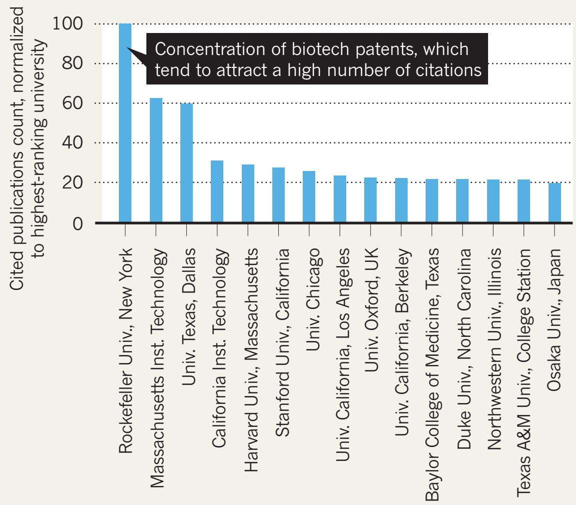
## INDUSTRY INFUSION

Some R&D conducted at higher-education institutes is financed by companies, China has the highest proportion of industry-financed R&D. Data are latest available (2012 or 2013)[2].



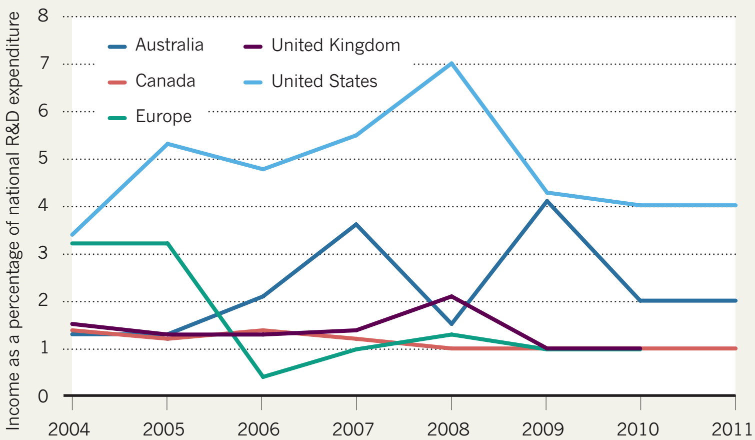

## R&D POWERHOUSES

US universities have the biggest impact on the development of breakthrough technology[3] (measured as publications cited by the most highly cited patents — those in the top 10% of patents cited by other patents).



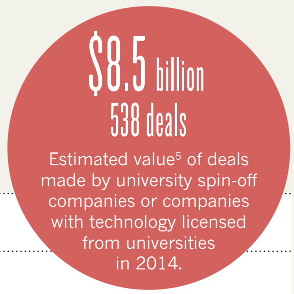

## LICENSING TRENDS

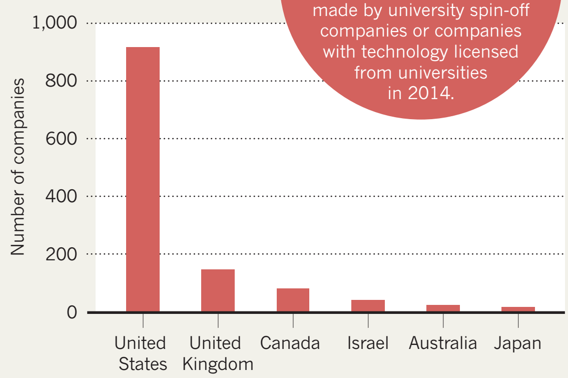
Licensing income from public research fluctuated during the 2007–09 recession, but remained generally stable between 2004 and 2011. However, only a handful of universities are responsible for the bulk of licensing in each country[2].




## SPUN OFF

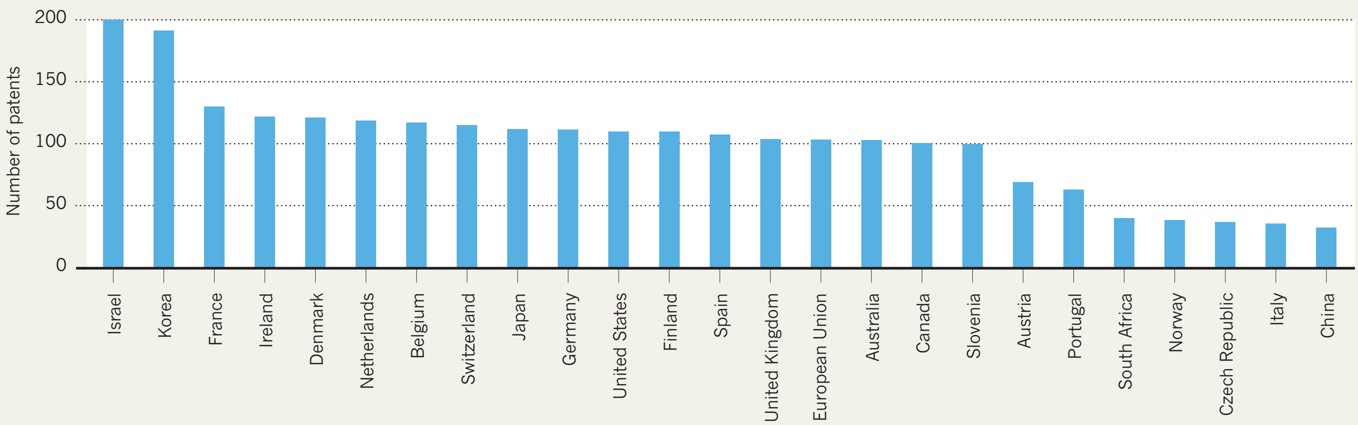
The total number of start-up companies formed from research institutions, including medical centres, in selected countries with a research-commercialization focus. Data are latest available (2013 or 2014)[4].




## PATENT LANDSCAPE

In 2014, Israeli public-research institutions filed the highest number of international patent applications, normalized by gross domestic product. How well this metric reflects commercialization, however, varies by country. Inventors may target their patent applications to specific regions and don't always file an international application[2].




Sources: 1. Innovation Policy Platform; 2. Organisation for Economic Co-operation and Development; 3. Leiden Univ. Center for Science and Technology Studies; 4. Assoc. Univ. Technology Managers, Australian Department of Industry and Science, UK National Centre for Universities and Businesses, Israel Central Bureau of Statistics, Japan Univ. Network of Innovation and Tech Transfer; 5. Global Univ. Venturing.


**Budding entrepreneurs at Harvard University's i-lab learn the skills needed to develop a new business.**

HARVARD INNOVATION LABS/EVGENIA ELISEEVA

START-UPS

# A sense of enterprise

*Universities aid entrepreneurs by helping them to turn their research into companies. In return, universities can reap financial benefits.*

**BY NEIL SAVAGE**

Michael Schrader knew he wanted to create a company, but he wasn't sure what it should do. After six years as a mechanical engineer in the automotive industry building plastic parts, in 2010 he began a master's degree in business administration at Harvard Business School in Boston, Massachusetts. In his quest for inspiration, he took a course in commercializing science at the Harvard Innovation Lab (i-lab).

The class heard presentations from researchers who among them had developed 17 different technologies that they thought had commercial value. One in particular caught Schrader's attention — a method devised by two engineers from Tufts University that uses a silk protein to stabilize vaccines. The vaccines could be formulated as powders and mixed with water when it was time to inject them, or embedded into a film that dissolves on the tongue like a breath-freshening strip. And, because they would not need to be refrigerated, they would be easier than conventional vaccines to distribute in places such as sub-Saharan Africa.

Along with other members of his class — an economics master's student, a former physics student earning a law degree and a postdoc in the chemistry department — Schrader spent the next few months looking into potential markets for the technology, making connections with business mentors and investors, and putting together a business plan. In 2012, the team founded Vaxess Technologies, which is attempting to bring vaccine formulations to market.

"We probably are a perfect model for how universities can forge together entrepreneurs and technologies to create companies," says Schrader, now chief executive of Vaxess. The technology has not yet entered clinical testing, but the company has raised more than US$5 million, hired 11 employees, and started filing patents of its own in addition to those it licensed from Tufts University.

Although universities often license technology developed in their research laboratories to existing companies that are looking for new products, they also move discoveries off the bench and into the real world by encouraging inventors to start businesses from scratch. They offer classes in entrepreneurship, introduce researchers to investors and business experts, and even launch their own venture-capital funds. The path is trickier for life-sciences spin-offs, which take more time and money to get off the ground, than for companies based on software or electronics. And Europe has not caught up with the United States in its ability to create businesses.

But universities are banking on entrepreneurs turning some of their research into products (see 'Start-up sampler).

**HUBS OF INNOVATION**

Universities tend to see commercialization as part of their remit to create and disseminate knowledge. "We exist on taxpayer money. We have an obligation to try to get our research out into society," says Regis Kelly, director of the California Institute for Quantitative Biosciences known as QB3. The institute is a collaboration between the Berkeley, Santa Cruz and San Francisco campuses of the University of California. It supports life-sciences research across the campuses and tries to bring that research to market by partnering with industry and promoting entrepreneurship.

Part of the mission of the University of Colorado Boulder's BioFrontiers Institute is to aid students and faculty members who want to start new companies, says Jana Watson-Capps, associate director of the institute. "It fits with what we want to do in providing an education for our students so that they can find jobs and be good at those jobs," she says.

A similar attitude is common in the United Kingdom. "We think it's important here in Oxford to see that the fruits of our research are actually developed to benefit society," says Linda Naylor, managing director of Isis Innovation, a company created by the University of Oxford to commercialize its research.

Harvard's i-lab, which was opened in late 2011 to help students in any of the university's schools to develop businesses, is a relatively new entry in a long line of such efforts at many academic institutions. Students learn about idea generation, business-plan development and marketing. Budding entrepreneurs can attend workshops on specific hurdles that they are likely to encounter, such as how to apply for a Small Business Innovation Research grant from the federal government. A group of 'experts in residence' provides students with business expertise and introduces them to potential investors. The i-lab holds competitions such as the President's Challenge, which awards ideas that address the world's big problems. Vaxess took the challenge's top prize of $70,000 in 2012, as well as winning $25,000 in Harvard's Business Plan Contest the same year.

Because the main thrust of the i-lab is education, the university never takes a stake in any of the companies created there, says managing director Jodi Goldstein. Any intellectual property developed in a Harvard research lab belongs to the university and must be licensed, but ideas generated in the i-lab belong to the students. Goldstein hopes that the i-lab can help a future Mark Zuckerberg or Bill Gates to pursue their billion-dollar idea while still completing their degree. "We have several pretty famous dropouts around here, and I don't think that's necessary anymore," she says.

As well as education and expertise, the i-lab provides a workspace for fledging companies. Meeting rooms, computer workstations and private storage space are available, as are a workshop for building prototypes and a pair of 3D printers. The i-lab is also planning to address one of the stumbling blocks that often trips up biology-based companies: finding a space to turn a discovery made in a university lab into a more marketable version. It is building a 1,400-square-metre wet lab with 36 research benches. When Vaxess reached that stage, it moved to LabCentral in Cambridge, Massachusetts. The provider of office and laboratory space takes care of regulatory requirements and provides administrative support and laboratory personnel so that new companies don't have to spend time and money setting up their own space. It opened in 2013 with a $5-million grant from the Massachusetts government (part of an initiative to bolster life-sciences business in the state) along with support from the Massachusetts Institute of Technology and the venture-capital arm of health-care giant Johnson & Johnson. Schrader considers this industry–government–academia web of support essential to his company's launch. "We have really taken advantage of this growing entrepreneurial ecosystem," he says.

*"We exist on taxpayer money. We have an obligation to try to get our research out into society."*

At QB3 in California, start-ups can rent lab space for as little as $85–100 per square metre per month. Unlike conventional landlords, who prefer to rent out an entire space, start-ups can rent a few hours in a fume cupboard or a shelf in a freezer, for example. "You only pay for what you actually use," Kelly says. Charging is important, mainly because it is a way of weaning its users off the university teat. "It gets people more used to being in the private sector," he says.

The need for lab space is just one reason why starting a life-sciences company can be much more challenging than, say, launching a business based on software. Any sort of pharmaceutical or medical device is subject to regulatory requirements, which leads to safety tests and clinical trials "If you're going to make a new drug you might need ten years and a billion dollars," says Watson-Capps.

These time and capital requirements make it much more difficult to drum up investment for a life-sciences start-up. Although investors might be willing to risk a couple of hundred thousand dollars on a promising software idea, most life-sciences companies need initial funding of a few million dollars. "Obviously, people don't want to throw away a million dollars, so they have to do a lot more due diligence," Kelly says. And because the time to realize a return on the investment can be so long, trading equity in the company in exchange for, say, legal services is not as popular as it is for other types of start-ups, he adds. These disparities are apparent in the investment statistics. Of the $77.3 billion in venture capital invested in the

## LICENSING TECHNOLOGY
### *Innovation income*

When it comes to commercializing research, universities often emphasize their desire to spread their discoveries, but they also reap financial rewards from licensing technology and investing in spin-off companies. Isis Innovation, for instance, took in £24.6 million (US$34.9 million) in revenue in 2015, of which it returned £13.6 million to its founder Oxford University, UK, more than double 2014's £6.7 million. The university also earned more than £30 million in cash and stocks from the 2014 sale of the games and technology company NaturalMotion (in which it had a stake of about 9%) to Zynga in San Francisco, California, for $527 million. NaturalMotion was co-founded in 2001 by Torstein Reil, then a PhD student in Oxford's zoology department studying neural systems. Reil used his research to create computer simulations that more accurately mimic how animals move, and turned them into a company that makes popular games such as Clumsy Ninja.

But licensing income tends to make up only a small part of a university's revenue stream. Harvard University in Cambridge, Massachusetts, which last year issued 50 licenses to patents it owns and saw 14 firms started on the basis of its technology, had licensing revenue of $16.1 million in 2015. But that is a fraction of Harvard's 2015 budget of nearly $4.5 billion, of which the university spent $876 million on research.

Jana Watson-Capps, associate director of the University of Colorado Boulder's BioFrontiers Institute, says that income from all licensing — not just from spin-off companies — is valuable to the university and goes back into funding research. However, she adds, licensing income is relatively small and comes so long after the initial investment that it's not a major consideration at the institute. A similar attitude prevails at Oxford. Although the university welcomes the licensing income, it's not the only motive for promoting spin-offs, says Linda Naylor, managing director of Isis Innovation. "The university is very clear it wants to create impact," she says. "They're not there to make any quick money." **N.S.**

**START-UP SAMPLER**
Universities seeking to commercialize research spin off scores of companies. These examples show the range of entrepreneurship spawned in the life sciences.

| Company | University | Technology | Founded | Financial milestone |
|---|---|---|---|---|
| OxSyBio | University of Oxford, UK | 3D printing of tissues for research and clinical applications | 2014 | £1 million (US$1.42 million) funding, April 2014 |
| Semma Therapeutics | Harvard University, Cambridge, Massachusetts | Creation of insulin-producing cells from stem cells to treat people with diabetes | 2015 | $44 million funding, March 2015 |
| Click Nucleic Acids | University of Colorado, Boulder | DNA analogues for applications such as therapeutics and biosensing | 2015 | Seeking funding |
| Zephyrus Biosciences | University of California, Berkeley | Tools to allow single-cell sequencing using western blot protein analysis | 2013 | $1.86 million funding, August 2014 |
| Clyde Biosciences | University of Glasgow, UK | Combination of stem cells and optical detection technology to test drugs for cardiotoxicity | 2012 | £2 million funding, April 2015 |
| Ex Scientia | University of Dundee, UK | Small-molecule drugs that bind to a combination of targets | 2012 | Pharmaceutical contracts worth $6 million |

United States in 2015, software companies took in $31.2 billion — 40% of the total. Pharmaceuticals and biotechnology received a mere 12%.

### PLAYING CATCH UP

Europe lags behind the United States in producing start-ups of any kind, but the situation is improving. "We're certainly seeing a lot more spin-outs than we were a few years ago," says Naylor. "There is more money around that is willing to go into the early stage."

She attributes that growth, in part, to the UK government's creation of the Seed Enterprise Investment Scheme in 2012, which provides tax breaks to investors in start-up companies. "The UK has been one of the leaders in providing tax incentives for investors in start-ups of all types," says Karen Wilson, who studies entrepreneurship and innovation at Bruegel, an economic think tank in Brussels. Other countries across Europe, as well as Australia, have created their own tax incentives for investors modelled on the British scheme, although Wilson says that they're often controversial, derided as tax breaks for the wealthy. In the United States, tax incentives vary by state. The biggest legal change in the United States to promote spin-offs came in 1980, Wilson says, with the passage of the Bayh–Dole act, which allowed researchers to profit from inventions created with federal funding.

US and UK Universities have even been creating their own venture funds in recent years to invest in their spin-offs. The University of Cambridge, UK, created Cambridge Innovation Capital in 2013 with an initial fund of £50 million ($71 million). In 2014, the University of California began a $250-million fund. In May 2015, Isis launched Oxford Sciences Innovation to raise an initial £300 million from investors. And, in January, University College London opened the £50 million UCL Technology Fund, and the University of Bristol, UK, started its own enterprise fund (see 'Innovation income').

Entrepreneurial ecosystems in which inventors can find facilities, investors and business experts to help them to launch their companies are important for creating successful spin-offs, and they've been growing around many European universities, Wilson says. "There are an increasing number of these entrepreneurial hubs that are emerging across Europe, which are spawning these innovative high-growth firms," she says.

In the United Kingdom, Cambridge is popular for life-sciences start-ups, and in Munich, Germany, the focus is mobile technology. In Switzerland, start-ups are clustered around the University of Zurich and the Swiss Federal Institute of Technology in Lausanne, where they focus on computing and technology. In Finland, Espoo is a hub: in 2010, three institutions combined to form Aalto University, which has strengths in communications, energy and design. Linked by a bridge across the Øresund strait, Copenhagen and Malmo in Sweden, make up another life-sciences centre. In the past year, however, the influx of refugees from the Middle East has led to a tightening of border security and made crossing the bridge more difficult for everyone.

The clampdown on migration within Europe, says Wilson, is making it harder for fledging companies to grow and spread. Expansion of their markets has always been challenging for start-ups in Europe, she says, where pushing into another country means dealing with differences not only in language and culture but also in taxes and other regulations. Many European companies get to a point at which, when they need to grow into a bigger market, they move to the United States, either of their own accord or at the insistence of their investors. "If you have a successful start-up in Italy it's much easier to go scale it in the US than it is to try to scale it across Europe," Wilson says.

But many life-sciences companies won't grow on their own, particularly if their innovation is a drug — their endgame is often to be acquired by a large pharmaceutical company once they have advanced their therapy to a promising stage.

Although life-sciences companies demand more resources than other types of start-up, they have one characteristic that can make them uniquely appealing to investors — the potential for curing a disease or improving human health. As Kelly points out, "Almost any rich person has a sick relative." If investors are going to risk their money, knowing that many of the companies they invest in will fail, they may prefer investments that have a potential for making a difference, he says. "If they're going to lose money on a business, they might as well lose it on something that could have some benefit to society." ■

PATRICK HO/VAXESS

Vaxess Technologies are using silk proteins (L), which are extracted from cocoons (R), to stabilize vaccines.

**Neil Savage** *is a freelance science writer based in Lowell, Massachusetts.*

PETER CANNING

Working with the Structural Genomics Consortium, researchers at the University of Oxford, UK, study new and often difficult to target proteins.

TECHNOLOGY TRANSFER

# The leap to industry

*The science done in university laboratories can change the world, but only when discoveries can be transformed into innovations.*

**BY JESSICA WAPNER**

The process of commercializing the discoveries made in university laboratories has come a long way over the past 30 years or so. "I didn't even know what that meant when I started out," says biomedical engineer David Kaplan at Tufts University in Medford, Massachusetts. Fifteen years and eight companies after his first patent, for a knee ligament made of silk, Kaplan is now well versed in the ways of technology-transfer offices (administrative infrastructure for ushering innovations out of the lab and into private development). The wisdom he has gained boils down to a few simple words: "It's an evolution," he says. And with shifting economic pressures, a drive to accelerate public access to innovations and changes to intellectual property law, technology transfer may be on the cusp of a major evolutionary leap.

Most historians agree that patent legislation originated in the Italian city of Venice in 1474. But for many centuries, universities in Europe and the United States were not involved in bringing new inventions to society. Because many universities were publicly funded, discoveries were published in the scientific literature, but were not patented. Industry and academia operated in vastly different spheres.

Licensing of inventions by academics became more prevalent in the early twentieth century. US chemist Frederick Cottrell received a patent for his device to reduce industrial pollution — an electrostatic precipitator — in 1908. The University of Wisconsin–Madison founded its technology-transfer office in 1925 to disseminate biochemist Harry Steenbock's discovery that irradiating food to increase vitamin D could treat rickets. Steenbock paid his own patent fees of US$300 (equivalent to roughly $4,000 today). When Quaker Oats offered him $1 million for his invention, Steenbock worked with university administrators to create an office that would allow the academic institution to benefit financially. The office licensed Steenbock's technology to Quaker Oats in 1927, leading to the introduction of breakfast cereal fortified with vitamin D. Bodies such as the US National Science Foundation, established in 1950, and the German Research Foundation, founded in 1951, increased government funding for academic research, but legislation allowing the commercialization of discoveries did not keep pace. Discoveries made by scientists through publicly funded research grants became the property of the governments that provided the money.

➲ NATURE.COM
To read more about the CRISPR–Cas9 battle, see go.nature.com/96ddzw

A few pathways from public invention to private commercialization did exist. The UK established the National Research Development Corporation (NRDC) in 1948 — a government body that led to innovations such as the first hovercraft in the late 1950s. In the United States,

private companies could enter into institutional patent agreements with universities, but it was a fraught process, with rules varying among universities and government agencies. By the late 1970s, of the estimated 30,000 patents accrued by the US government through federally funded research, only around 1,200 were licensed and even fewer had made it to market. In Europe, legislation was mostly lacking. Germany's Employees' Inventions Act of 1957 gave more autonomy to academic inventors, but in general there was little interest across Europe in commercializing publicly funded research.

In the United States, the 1980 Bayh–Dole Act catalysed a surge of interest in commercializing academic research. The landmark legislation continues to provide a legal framework for patenting discoveries made using federal grant money. In the United Kingdom, the biggest shift came in 1985, when the government eliminated the monopoly that the British Technology Group, a public body, had on commercializing publicly funded innovations — a move that was followed by an increase in academic entrepreneurship. Several other European countries, including Germany, Denmark and Belgium, also have technology-transfer legislation, but laws governing this practice vary widely. Some are more restrictive on individuals, allowing universities to retain ownership of an invention instead. Others permit inventors to own patents derived from publicly funded research. This variation led the international group the Organisation for Economic Co-operation and Development to consider whether a Bayh–Dole-type policy should be adopted by the organization's member countries.

Since Bayh–Dole was enacted, technology-transfer offices have proliferated at universities in the United States and elsewhere. In 2014, at least 6,300 licences were secured by technology-transfer offices in the United States. Technology transfer has made available discoveries such as cancer drugs, recombinant DNA, imaging diagnostics and nanotechnology — in the United States alone, more than 23,000 patents have been filed by universities.

But technology transfer is facing several challenges. In the United States, which is the largest generator of academic innovations, federal grant budgets have shrunk or at best remained flat since 2003. In the United Kingdom, despite some capital investment, the budget for basic-science research has remained at £4.7 billion (US$6.7 billion) annually for the past 6 years. And how changes to the US patent system will impact commercialization is unknown — the United States has adopted a first-inventor-to-file rather than a first-to-invent structure, initiated by the 2011 Leahy–Smith America Invents Act, bringing it more in line with the rest of the world.

The first-to-invent system awards patents to the individual who first conceived the idea, created a workable prototype and then filed a patent. The first-to-file approach awards the patent to whoever submits the paperwork first, regardless of when the idea was conceived. The change to the US system may reduce interference proceedings — lengthy and costly battles that follow claims to a patent by separate parties, as is currently happening between the Broad Institute of MIT and Harvard and the University of California, Berkeley, over CRISPR–Cas9 gene-editing technology. However, the first-to-file approach could shift the focus away from carefully ensuring that an innovation is workable in favour of racing to file paperwork on an incomplete idea. The change could also favour large companies — with the resources, such as staff and attorneys, to handle large volumes of patents — over smaller companies or independent inventors.

"We're at another inflection point," says John Swartley, executive director of the Penn Center for Innovation (PCI), the technology-transfer office at the University of Pennsylvania in Philadelphia. To face these challenges, technology-transfer offices need to find new ways to work with private companies, scientists and outside investors, while maintaining their own integrity. "We can never forget that we are, at core, an academic institution," says Swartley.

## COPING STRATEGIES

One of the most difficult aspects of moving a technology from academic concept to valuable product is crossing the chasm between early innovation and readiness for licensing — a stretch often referred to as the 'valley of death'. "One of the greatest challenges for academic technology transfer is trying to interest either established companies or venture investors in our early-stage discoveries," says Fred Reinhart senior adviser in the technology-transfer office at the University of Massachusetts Amherst.

High-volume crystallization plates used by the Structural Genomics Consortium.

ELLA HUGHMAN

"Almost all of them have gone to later-stage, less risky investments."

That hurdle also exists outside the United States. "There's always been a relative shortage of cash at this early stage," says Steven Schooling, director of engineering and physical sciences at University College of London Business (UCLB), the technology-transfer office at UCL.

Many universities are providing internal funding to bridge the valley of death, along with seed funding for even earlier stages of research when no other grant support exists. At North Carolina State University in Raleigh, the Chancellor's Innovation Fund provides awards of up to $75,000 to researchers whose work has garnered encouraging feedback from an outside company. "It's not huge money," acknowledges Kelly Sexton, director of the office of technology transfer at the university. But the amount is enough to help academics through the proof-of-concept stage. "There's kind of a sweet spot where this can be useful," says Sexton. In January 2016, UCL launched a £50-million UCL Technology Fund, which can be used to support researchers through the proof-of-concept stage. The money is provided by the European Investment Fund (EIF) and technology-commercialization company Imperial Innovations, and will be managed by the venture-capital firm Albion Ventures, which is also a contributor. The aim is to overcome the challenge of attracting and sustaining interest from investors who generally have to wait a long time to see a return. To make the long-term investment more attractive, the fund will pay out an annuity over 15–20 years — an approach that may avoid the drop-off that is frequently seen with the conventional venture-capital model of raising capital in multiple rounds with the hope of reaping benefits from a trade sale or initial public offering. UCLB and Albion decide which researchers receive the funds, but follow strict return-on-investment criteria set by the EIF. "This isn't charity money," says Schooling, "and that means we have to be selective."

Charitable foundations that focus on a single disease are also becoming an increasingly prominent piece of the tech-transfer puzzle — a variety of venture philanthropy (see page S43). The approach has already led to several drug licences. For example, an experimental treatment for multiple myeloma, ricolinostat, was created as a result of research at the Dana-Farber Cancer Institute in Boston, and the Broad Institute of MIT and Harvard in Cambridge, Massachusetts. The investigators formed Acetylon to develop the technology, and the US-based Leukemia and Lymphoma Society contributed $5 million towards the phase I clinical trial. US biotech firm Celgene subsequently invested $100 million in the development of ricolinostat, a payment that included an exclusive option to buy the licence from Acetylon. The drug is now in phase II trials for multiple myeloma. The Leukemia and Lymphoma Society have also partnered with

Celator Pharmaceuticals in Ewing, New Jersey, to speed up the study of the acute myeloid leukaemia drug CPX-35, including an initial $4.1 million for the phase II study followed by an additional $5 million for the phase III trial.

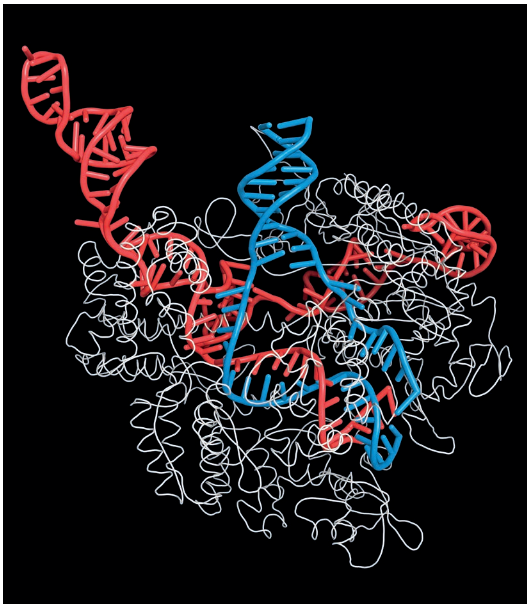
Some technology-transfer offices are changing their entire approach to working with private companies. At the Penn Centre for Innovation, the focus is shifting towards cultivating a few strong business relationships, rather than cold-calling hundreds of companies for every invention, says Swartley. At some institutions, pharmaceutical companies enter into research agreements with a specific laboratory or investigator, this offers a more collaborative approach to academic research.

The trend towards a more "holistic relationship", as Reinhart puts it, is allowing technology-transfer offices to avoid investing too much time in specific deals. Focusing instead on a long-term relationship between universities and companies enables "a better understanding of mutual needs", says Reinhart. "That's what the smart universities are doing these days." Reinhart cites the Office of Industry Engagement at Georgia Institute of Technology in Atlanta, the Office of Innovation and Industry Engagement at Michigan Technological University, and the Office of Technology Commercialization at Purdue University in West Lafayette, Indiana, as examples of technology-transfer offices moving towards this approach.

Start-up companies launched by principal investigators are becoming increasingly common, particularly when large companies are unwilling to assume the risk, even after the proof-of-concept stage. Building a successful product through a spin-off company can lead to lucrative deals later and allow the original researcher to continue working largely autonomously. In the United States, 818 start-up companies were formed on the basis of academic patents in 2013, a 16% increase from 2012 (L. Pressman *et al. The Economic Contribution of University/Nonprofit Inventions in the United States:1996-2013*; BIO, 2015).

*"We can never forget that we are, at core, an academic institution."*

The increase partly stems from universities being seen as sources of innovation and job creation, not merely sheltered places of "learning, teaching, and getting degrees", says Schooling. "We've moved beyond that." According to the US Association of University Technology Managers, 4,000 start-up companies in the United States have formed as a result of university innovations since 1980, and these have led to 3 million jobs. Schooling recalls that in the early 1990s, when he and fellow researchers founded a spin-off for work they had done at Manchester Metropolitan University, UK, his group was viewed as "slightly failed academics". Colleagues questioned their commercial activity. "Nowadays its part of how university academics are assessed," he says.



MOLEKUUL/SPL

Model of the CRISPR–Cas9 gene-editing complex.

## MORE SERIOUS REPAIR

Although these approaches are changing how technology-transfer offices operate, some researchers see the need for a more severe overhaul. For medical advances in particular, profit-driven privacy and competition spurred by the licensing infrastructure may be obstructing progress. "The current way we're doing drug discovery is too costly, too risky and too slow," says Chas Bountra, a member of the Structural Genomics Consortium (SGC) at the University of Oxford, UK. "The whole process is incredibly inefficient." Companies duplicate efforts and a large proportion of the compounds developed fail to show a benefit in clinical trials. Most troubling of all, he says, is that patients are sometimes treated with experimental medications that would already have been shelved, if data were shared earlier and more openly. "It's a horrendous waste of money, a waste of people's careers, and a waste of patients' willingness to participate in research," says Bountra.

As part of the SGC, Bountra is taking a radically different approach to therapeutic innovation. The consortium receives funding from several pharmaceutical companies, charities and government organizations. The large collection of funders means that resources are pooled and risk is shared, so that no single investor is shouldering the burden of early-stage development. Research is focused solely on novel proteins — often substances that have been deemed impossible to target. The tools developed to generate a potential drug are then made freely available. Data from preclinical studies are published immediately. "We tell the whole world about it," says Bountra.

He is not alone in encouraging the open-innovation approach. The Harvard Stem Cell Institute and the Biodesign programme at Stanford University, California, for example, are also taking steps towards a more open approach. David Brindley, who studies health-care translation at the Centre for the Advancement of Sustainable Medical Innovation (a partnership between Oxford and UCL) contends that the translation of technology from lab to bedside has been slowed by "disincentives for people along the chain to communicate and work together effectively".

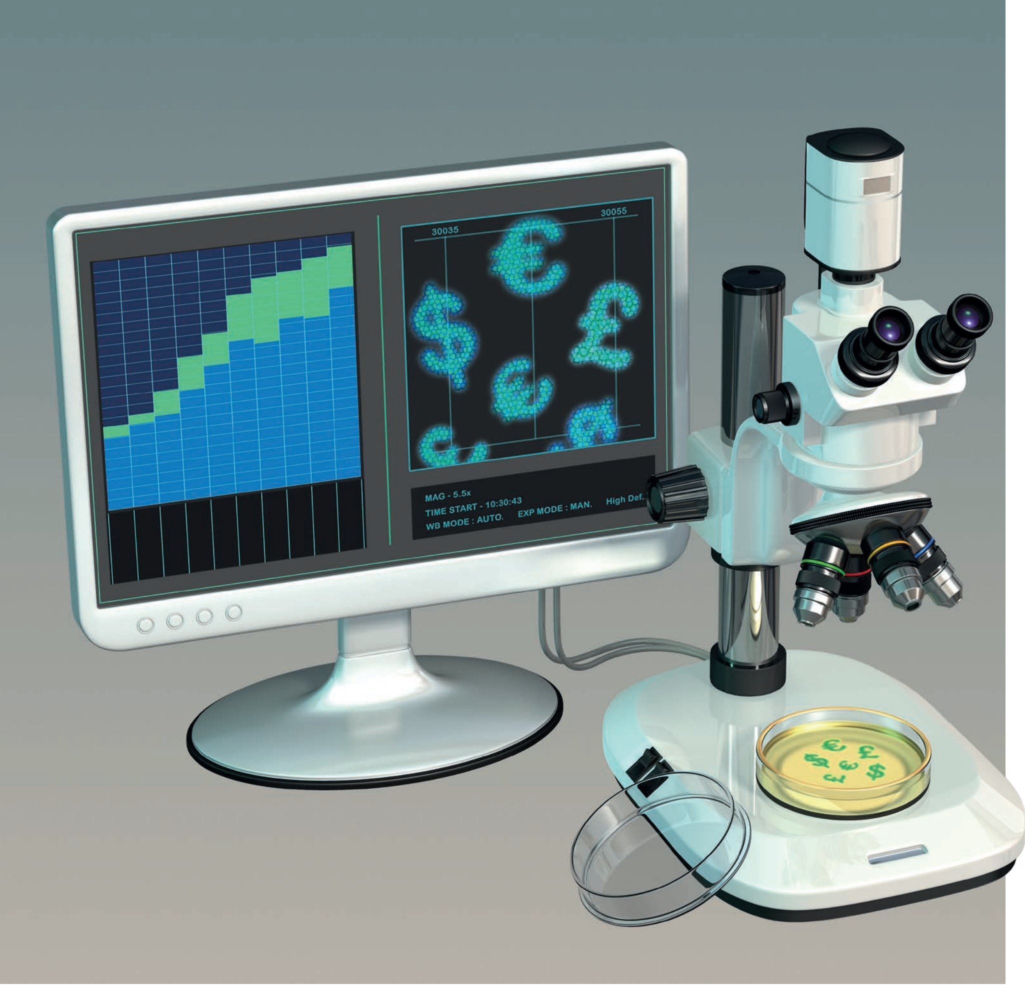
Brindley says that changes that better align the interests of academia with industry would help. Tenure applications, for example, could take entrepreneurial activities into consideration. He also advocates altering the conventional financial arrangements that surround university-born innovation. "Academia shouldn't expect industry to pay huge licensing revenue for research they funded in the first place," says Brindley, "and industry needs to be more reasonable in their expectations of research timelines."

Whatever route technology-transfer offices take, the most important need is to stay flexible, particularly in light of the increasing number of gene-based discoveries that raise ethical and proprietary questions that may not have been accounted for when the Bayh-Dole Act was passed. Who owns a gene? Can a gene be owned? The current legal battle over the patent for CRISPR–Cas9 may have a considerable impact on scientific innovation. The gene-editing technique is allowing all manner of genome alterations that could bring huge benefits, such as cures for disease and pest-resistant crops.

Although the financial stakes for the opposing parties are high, the broader relevance of the case may be minimal. Interference proceedings are connected with the first-to-invent patent system, and so when the ruling is made, it may not carry much weight in the first-to-file era. Still, the case could have broader ramifications on university-driven innovation, potentially forcing the creation of new legal frameworks for gene-based discoveries, such as the right to patent these innovations or to specify what can be done with them. Whether the quest to fill personal and university coffers will delay broader distribution of the lifesaving fruits of taxpayer-funded research remains unclear.

The CRISPR–Cas9 controversy stands in stark contrast to the lack of financial incentives favoured by those behind the SGC. As Bountra sees it, that transparency, and the academic freedom that it provides, is paramount to ensure that novel, effective medicines reach people as quickly as possible. "Tomorrow is too late," he says. "They want them today." ■

**Jessica Wapner**, *a freelance writer in Brooklyn, New York, is the author of* The Philadelphia Chromosome.




JOHN HARWOOD

ASSESSMENT

# Academic return

*A broader understanding of 'impact' could help governments to measure the diverse benefits of their investment in research.*

**BY MICHAEL EISENSTEIN**

When Julia Lane began working in scientific-funding policy she was quickly taken aback by how unscientific the discipline was compared with the rigorous processes she was used to in the labour-economics sector, "It was a relatively weak and marginalized field," says Lane, an economist at New York University.

In 2005, John Marburger, science adviser to then-President George W. Bush, felt much the same. He called on researchers and policymakers to focus on the "science of science policy", an empirical assessment of outcomes and returns from funding agencies such as the National Institutes of Health (NIH) and National Science Foundation (NSF). "When the Congressional Budget Office does simulations of the effects of investment in areas like tax or education policy, they have models and processes," says Lane. "But he said that when it comes to science, essentially all we say is 'send more money.'"

Around the same time, the UK government also began to explore how to significantly increase the economic impact of the country's research and development (R&D) investments. According to Lane, such efforts have historically been a low priority, because R&D accounts for only a small percentage of the economy — typically less than 3% of the gross domestic product (GDP), mostly from the private sector. However, public funding of basic research still represents a considerable sum.

In 2013, the United States spent more than US$40 billion on research at university- or government-run laboratories. Finding out what comes of this expenditure is crucial for economic reasons, but also has a moral dimension. "We can't sit in an ivory tower and expect the taxpayer to pay our salaries and not ask any questions," says Ben Martin, who specializes in science and technology policy at the University of Sussex, near Brighton, UK. Over the past 10–15 years, economists and policy experts have been trying to build smarter tools to answer such question about how public research investments pay off — a process that has entailed an examination of what precisely it means to get a return on R&D.

### NUMBER CRUNCH

The earliest efforts approached this question purely in economic terms. Martin and his colleague Ammon Salter, now at the University of Bath, UK, reviewed[1] studies on the benefits of publicly funded basic research — including pioneering work by the US economist Edwin Mansfield, who surveyed businesses to learn what proportion of their products arose from this type of research and determined a 28% rate of return. However, they found that these studies generally took an overly simple approach to tackling a complex question. "We concluded that there are too many conceptual, methodological and empirical problems with these kinds of efforts," says Martin.

Economic analysis is complicated by numerous intermediate indicators of performance (number of patents licensed, for example), as well as more direct impacts such as the number of products sold. The true impact emerges from a combination of these factors. "The temptation to come up with a number for an impressive-looking economic return can be strong," says Adam Jaffe, director of Motu Economic and Public Policy Research in Wellington, New Zealand, "but I'd argue that you should look at a range of different indicators, including qualitative information."

The most comprehensive studies tend to be technology- or field-specific. In 2008, the research institute RAND Europe teamed up with academics to analyse the impact of UK research grants for cardiovascular disease and stroke[2]. They used a strategy called the payback framework, which combines surveys and data analysis to assess the impact of research across many domains, rather than just basic economic gain. "You might prove that a method of developing stents for heart disease has generated jobs in industry, new skills, new research areas, benefits for patients who receive stents, and economic benefits in terms of helping these patients to return to work," explains Steven Wooding, a researcher at RAND. "Then, at the other end, you can figure out what each one is worth." They concluded that every £1 (US$1.43) invested in cardiovascular-disease research between 1975 and 1992 generated £1.39 of return in economic and health terms. However, this method is labour intensive and designed for biomedical research.

Patents based on academic research can provide a useful general indicator of commercial interest in a particular invention. But this is not always straightforward to interpret because not all patents become products. Furthermore, the public-sector origins of private-sector patents are not always obvious. A team led by Danielle Li at Harvard Business School in Boston, Massachusetts, has attempted to clarify these links by forging connections between NIH grants,

the papers that they generate and patents citing those papers[3]. "She's used that to see, for example, whether NIH funding in a given therapeutic area advances the treatment options in that area," says Jaffe. "That's getting a little closer to real impact."

Such analyses depend on well-organized data. In 2009, the UK Medical Research Council (MRC) began using software called Researchfish to collect relevant information on the productivity of its researchers, including articles, patents and spin-off companies that arise from a grant. This programme has since expanded to encompass all of the UK Research Councils as well as other funding agencies; Ian Viney, director of strategic evaluation and impact at the MRC, anticipates that more than 40,000 UK researchers will file these reports in 2016.

In the United States, the Institute for Research on Innovation and Science (IRIS) relies on a more automated approach, drawing data directly from participating research universities. IRIS is a descendent of a federal programme created by Lane and colleagues at NIH and NSF to track research jobs created by President Barack Obama's 2009 economic stimulus, which included $52 billion for R&D. According to executive director Jason Owen-Smith, a sociologist at the University of Michigan in Ann Arbor, IRIS has already partnered with 24 universities, representing $15 billion of R&D funding. "Our goal is to involve every institution that gets at least $100 million of federal R&D, as well as flagship state and land-grant universities," he says — a scope that would include data on more than 90% of all federally funded R&D.

The premise of the US assessment efforts is that scientists themselves — rather than the publications or patents — are the main vehicles by which research fuels economic growth. Owen-Smith says that, in his experience of university technology-transfer offices, such organizations generally believe that "disembodied inventions aren't particularly valuable", and that for real economic pay off "you have to have a member of the original research team involved in the commercialization." IRIS data allow observational experiments that can directly test this people-centric model by tracking how scientific training affects career trajectories and returns to industry. Preliminary IRIS data indicate, for example, that a science doctorate improves a person's chances of entering a high-tech industry, which will result in higher wages and greater productivity.

## BEYOND PROFIT AND LOSS

Disentangling causation from correlation remains difficult. "You can look at the impact on particular researchers who were funded compared to those who weren't," says Jaffe, "but that's not quite the same as asking how a world that has a 'war on cancer' differs from one that doesn't." Large-scale data collection programmes such as IRIS and Researchfish could clarify this by examining the changes associated

TESSA SHAW/JULIA LANE

Julia Lane (centre) explains data-collection tool IRIS.

with an influx of targeted spending such as the NIH Precision Medicine Initiative.

The long time lag between inception and commercialization can also be a major confounder. "People tend to use at least 20-year time windows," says Robert Tijssen, chair of science and innovation studies at Leiden University in the Netherlands. "You can't expect any economic impact in the narrow sense from a research programme within two or three years — that's only the case for exceptional research breakthroughs." Wooding and colleagues have noted that many independent analyses have described a consistent gap of 17 years from initial publication to economic impact across biomedical fields, whether that impact represented formal adoption of a medical intervention or marketing of a new drug[4], although the nature of these lags remains poorly defined.

> *"Our goal is to involve every institution that gets at least $100 million of federal R&D."*

Money isn't everything. Many research outcomes can benefit the economy more indirectly through factors such as environmental sustainability or improved quality of life. The United Kingdom has taken the lead in comprehensively measuring this diversity of benefits with its Research Excellence Framework (REF). REF, which helps to determine the allocation of funding to individual universities, relies on peer-reviewed case studies submitted by each institution that offer insight into both research 'quality' (in terms of outputs such as published papers) as well as impact on areas that range from the economy and health to public policy and culture. For example, the impact of medical research might be measured on the basis of evidence of public debate or changes in clinical or public-health guidelines. Viney notes that the first iteration of REF, completed in 2014, reflected a huge variety of impacts: "There's hardly any walk of life or part of society that research doesn't have some bearing upon."

But REF is labour-intensive and Martin is concerned that future iterations may become even more time-consuming and expensive. "There is probably an optimum point beyond which the costs become greater than the benefits, and we're not very good at working out what that optimum point is," he says. Nevertheless, the concept of impact assessment is being emulated in other countries, including the Netherlands, Norway and Australia (see page S22). Meanwhile, researchers developing IRIS and Researchfish are exploring strategies to track these impacts in a more automated and structured way; for example, by tracking citations in government policy statements.

## EMPIRICAL EVIDENCE

The surge in interest could transform research assessment into a thriving, evidence-based subfield of economics. With hard numbers to hand, research funders and university administrators could gain the tools for making decisions that were once largely guided by dogma or instinct, such as determining what are the most effective ways to inject funding into new fields. Metrics could also help policymakers to identify the optimal GDP percentage that a nation should be spending on R&D.

The extent to which policymakers will respond to such a multidimensional view of socio-economic impact will vary. For some governments, demands for a sound-bite-friendly number that reflects simple return of investment may prevail. In 2012, Jaffe was part of a working group for the US National Academy of Sciences, which looked at the various ways in which scientific impact can be measured, only to find that politicians were mostly interested in lists of economic winners and losers. "They wanted us to tell them, in effect, whether the rate of return in energy research is higher or lower than in biomedical research so we can figure out where to redirect money, and I think that's a fundamentally misdirected question," he says.

The economic assessment of science is an inevitability, says Owen-Smith. But if academics take the lead, they can strive to ensure that the assessment is fair, intellectually rigorous and a mechanism to grow, rather than constrain, the scientific endeavour. "We know as little about what our key social and economic needs will be 30 years from now as we might have known about the Internet in 1974," he says. "We should be managing our publicly funded R&D system as a capacity and infrastructure for our society to hedge against an uncertain future." ■

**Michael Eisenstein** *is a freelance science writer in Philadelphia, Pennsylvania.*


1. Salter, A. J. & Martin, B. R. *Res. Policy* **30,** 509–532 (2001).
2. Health Economics Research Group, Office of Health Economics & RAND Europe. *Medical Research: What's it Worth?* (UK Evaluation Forum, 2008).
3. Azoulay, P., Graff Zivin, J. S. Li, D. & Sampat, B. N. Public R&D NBER Working Paper 16-056 (2015).
4. Slote Morris, Z., Wooding, S. & Grant, J. *J. R. Soc. Med.* **104,** 510–520 (2011).

A researcher at the Australian Institute for Bioengineering and Nanotechnology works on a project run in collaboration with the public body Queensland Health.

AUSTRALIA

AUSTRALIAN INST. BIOENG. NANOTECHNOL., UQ

# Engagement upgrade

*The value that Australia places on publication quality over quantity has elevated it into the top echelon of science. Can it now improve its flagging track record in commercialization?*

BY BIANCA NOGRADY

It is often said that when it comes to research excellence, Australia punches well above its weight. Despite a population of only 23 million, the country ranked 12th in the global *Nature* Index (see go.nature.com/1dbcsr), which tracks the contributions of countries and institutions to high-quality scientific journals. This impressive performance can be partly attributed to a research-output measure introduced in 2010 to encourage quality over quantity. The Excellence in Research for Australia (ERA) metric looks at the breadth of research from universities and evaluates the quality against international standards. "The ERA exercise, focusing on quality of the outputs at universities, has been very beneficial to the university system in Australia," says Aidan Byrne, chief executive of the Australian Research Council in Canberra, which administers the framework. "It has been a focus that all of the universities in Australia positively responded to, and it added to the strength of the Australian university system."

But in the shadow of Australia's research performance lurks the country's poor track record for translating that research into economic impact. The 2015 Global Innovation Index (S. Dutta *et al.* (eds) *The Global Innovation Index 2015: Effective Innovation Policies for Development*; Cornell University, INSEAD & WIPO, 2015) ranked Australia 72 out of 141 countries for innovation efficiency. The country places respectably high — number 17 globally — on the overall innovation-index ranking, which takes into account factors such as regulatory environment, investment, education and general infrastructure.

"Our challenge now is to look at what is the next step," says Byrne, namely tailoring university research "to the benefit of the Australian commercial business and community more broadly".

## ENGAGEMENT MISSION

After more than a decade of discussion, analysis, pilot studies and initiatives thwarted by political change, Australia is embarking on a mission to bring its reputation for research commercialization into line with its track record for research quality. Buoyed by support from the federal government's National Innovation and Science Agenda, announced in late 2015, and recommendations from last year's review of research policy and funding, a multi-institution committee is developing a system to measure the amount of research engagement, interaction, knowledge transfer and collaboration between universities and potential public- and private-sector users of the research.

This process, called Research Engagement for Australia (REA), began in 2014, when the Australian Academy of Technology and Engineering (ATSE) began exploring ways to measure research engagement. The project's steering committee reviewed many options, but concluded that calculating the amount of money that the research attracted from the end user was the most suitable, says ATSE president and chair of the REA steering committee Peter Gray. "Dollars are auditable and they are a true measure of collaboration," he says. "It's a good independent measure of the degree of commitment by the end user to the collaborative research programme."

The measured income includes money from certain competitive grants, government contracts, industry contracts, funding from philanthropic groups, and money earned from participating in collaborative endeavours, such as one of the government's Cooperative Research Centres.

Income seems to be well suited to act as a measure of assessing commercialization. But if income is the numerator, what is the denominator? The committee initially considered three metrics with which the figure could be compared: full-time equivalent hours for that field, total national activity in that field and the

JOSE FUSTE RAGA/CORBIS

university's total operating income.

Gray says that the committee is leaning towards using just the latter two, acknowledging that full-time equivalent hours "are a pretty rubbery number". The advantage of measuring engagement in dollars is that all the necessary data are already collected and reported through the Higher Education Research Data Collection and the ERA programme.

To look at the resources that such a programme might demand, the ATSE ran a pilot of the metric with universities in Queensland and South Australia. Gray says that because the new programme requires only a handful more details than are routinely collected, it imposes little additional burden on the universities.

Sydney Harbour, Australia. The country excels at research, but its ability to commercialize this is lacking.

## BEYOND THE NUMBERS

Not everyone is satisfied that income alone is enough to demonstrate a university's research impact in a particular field. John Dewar, vice-chancellor at La Trobe University in Melbourne, and chair of the six-institution consortium Innovative Research Universities, says that there is a need for qualitative as well as quantitative assessment that will show not only income but also impact. The approach proposed by the ATSE measures the amount of money that research attracts from industry. "But we don't think that's the link of the chain that we need to improve," says Dewar. "We think it's the second link — taking ideas and innovations and making something useful." The consortium is therefore arguing for the inclusion of panel-based assessments of the value of university research for end users. "We don't see any alternative to some form of qualitative data where you talk to your industry partner and ask what impact has this had on your business or your sector of the economy," Dewar says. The consortium has suggested adopting a case-study-based method such as that used in the UK's Research Excellence Framework.

But this idea elicits a nervous reaction from some. In 2014, the assessment of 6,975 impact case studies at 154 UK universities cost £246 million (US$347 million). "The case-study approach is very, very expensive and time-consuming, and it's also very difficult to track outcomes back to either an individual or institution," says Margaret Sheil, provost at the University of Melbourne and a member of the REA steering committee.

The REA's originators are listening to both sides. Gray says that the ATSE is keen to avoid the case-study approach, but is open to a provision for qualitative data, particularly if an institution wants to highlight an especially fruitful engagement outcome. "From the pilot study, we thought we probably should give people the opportunity, if they've had a big success, to write a little vignette about why they have been successful," Gray says.

Another concern about focusing almost exclusively on income is how this will work for the humanities, arts and social sciences, for which engagement and impact might not be as easy to quantify as they are in science, technology, engineering and maths. "When you look at just about any indicator, there are very strong discipline variations," says Byrne. The Australian Research Council has been tasked with the development and piloting of the assessment in consultation with the research sector, and Byrne says that it intends to take an approach similar to that taken for the development of the ERA metric. The goal, he says, is "to get a sense of what are the most significant drivers for your discipline that will tell you something about engagement and impact".

The objective of the REA programme is for universities to value achievements in the commercialization of research excellence alongside publication successes. This is a similar goal to that of the ERA metric, which encouraged academics to focus on publication quantity as well as on quality by providing a regular, nationwide 'stock take' of universities' research strengths and weaknesses.

> *"Dollars are auditable and they are a true measure of collaboration."*

Evidence of the success of the ERA can be seen in the consistent improvement of Australia's research rankings since the framework's introduction. Some policymakers think that the REA can do the same for the other half of the innovation equation. The metric's impact will lie not in its influence on funding but, as for the ERA, on the message it sends to the academic sector about what the government values.

"What a measure like REA is designed to do is counterbalance at the institutional level; to say that we also need to ensure that we've got engagement happening," says Sheil. Although the intention is to change the institutional mindset, the hope is that this signal will be heard at all levels of academia, particularly among younger academics and students who might be more likely to contemplate investing their time and energy — and risk a gap in their publishing record — with a commercial endeavour. A measure like the REA could encourage universities to support a broader range of career trajectories, including commercialization and industrial collaboration, for their staff.

But Australia still faces the challenge of a relatively risk-averse and conservative commercial ecosystem, which lacks the kind of deep pockets found in other parts of the world, Sheil says. "We don't have a Silicon Valley where you can be an academic, go and try your spin-off, then come back to your university; we don't have the venture capital that's attracted by that," she says. Financial capital in Australia is tied up in property, mining and retirement funds, and the country has relatively few private investors. But if the REA programme can impel universities and academics to improve their engagement with industry, and translate research into commercial success despite these constraints, it could establish a model for many other countries that face similar challenges.

These are early days for the REA, but the momentum is strong. In its National Innovation and Science Agenda, the government singled out the need for a measure such as the REA to be part of a national assessment of university research performance.

Although the aim is for the framework to re-adjust the historical focus on publication record, there is a risk that too much emphasis will be placed on research commercialization, jeopardizing financing of fundamental research. But those involved in the development are determined not to risk Australia's track record in basic research, stressing that this will require deft compromises.

Perhaps a light touch will be enough. "Universities are very good at responding to even the smallest signal from government," Dewar says. "The signal being sent is of modest changes, but over time that could have a quite significant ripple effect across the sector." ■

**Bianca Nogrady** *is a freelance science writer in Sydney, Australia.*

# Q&A Horst Domdey

# Bavarian biotech

*After starting one of Germany's first biotech companies, biochemist Horst Domdey co-founded BioM, a non-profit organization that has managed and developed Munich's biotechnology cluster since 1997. He talks to* Nature *about nurturing the entrepreneurial spirit in "a country of competitions".*

ANDREAS BROECKEL

**When you co-founded BioM in 1997, what were you trying to accomplish?**
In those days, Munich was nowhere — just a tiny spot on the biotechnology map. There were just 30 companies, which were not well financed. But Munich had research capacity. Two elite universities, three biologically-oriented Max Planck Institutes, and the Helmholtz Centre Munich — it was, and is, a great place for excellent science. Together with Ronald Mertz from what is now known as the Bavarian State Ministry for Economic Affairs and Media, Energy and Technology, we entered Germany's BioRegio contest to become a model region for biotechnology, and we were one of the three winners. Using the prize money of €25 million (US$28 million), we developed a concept for how to turn the handful of companies into a bioregion. In the beginning, we got a lot of help from the state of Bavaria, as well as from the federal government. The government put some of the money that they had made from selling stocks in big companies into science and innovation initiatives. Most of it went into infrastructure such as scientific buildings, but some was also put into incubators.

**What was your strategy for creating a bioregion?**
First, we decided that the prize money would go only to biotech start-ups. The strengths of the Munich biotech cluster lay in drug development — an area where you really have the chance to become renowned. Second, we wanted to build a close network between all partners, which included the emerging biotech industry, the pharmaceutical industry, finance institutions, and universities and research institutes. Third, we wanted to address the challenge of how to turn a scientist into a business person. We combined the money provided by local and national government with money from the pharmaceutical industry and banks to generate something like a seed fund to finance start-up companies. Altogether, we funded more than 40 start-ups, half of which were successful. There are now 250 life-sciences companies in Munich, of which 120 are small- or medium-sized companies.

**What do you think is the best approach to commercializing academic research?**
Investors have realized that scientists are not the best business people. You need different management. We still identify the best science, but now we find a serial entrepreneur or an existing company who knows how to transform the research into a commercial product. The scientists still play a very important part: the companies rely on them as advisers. But unlike when I was in academia around 20 years ago, scientists don't have to leave the university to commercialize their research.

**How do you balance the needs of the different stakeholders in the cluster?**
The biggest question in starting a new company is who owns the intellectual property.

Universities do not have enough money to finance all the patent applications that their researchers want to file. The Technical University of Munich only files a patent if there is at least a letter of intent from a company that they will license it. That's easy in engineering, where vehicle manufacturers such as BMW and Audi are waiting in line. But it's not possible in the life-sciences area. There is a tendency in Germany for university tech-transfer offices to take a stake in the start-up of 20% or 30%. That doesn't leave the founders with that much ownership, which discourages them from proceeding with the start-up. Why should a scientist find time to write a business plan for something when he knows that it's difficult, if not impossible, to receive funding or financing for the business? That is concerning, and it doesn't make sense because German universities are taxpayer-funded and so do not have to make money on commercial enterprises. And, if the start-up is successful, the government receives a lot of money in tax revenue.

MEDIGENE AG

Deep frozen cells are removed from storage at German biotechnology company Medigene.

**How do you help companies in your cluster to succeed when there is limited venture capital?**
One of the most interesting scientific ideas we have seen in the past few years was a type of cancer immunotherapy related to dendritic and T cells, invented at the Helmholtz Centre. The idea received €500,000 through one of our competitions. We wrote a business plan, formed a company — Trianta Immunotherapies — and looked for investors, but didn't find any. So we decided to circumvent this problem by connecting the Helmholtz scientists to Medigene, a local company that I helped found. Trianta was merged into the established company, and the founder of the start-up became the chief scientific officer of Medigene. The company raised more than €60 million to invest in the start-up on the international capital markets thanks to already being listed on the stock exchange. Another company, CorImmun, which received funding through a national programme coordinated by BioM, developed a peptide that can be used to treat a severe form of heart failure. The peptide neutralizes an autoantibody that binds to a receptor in heart cells, causing the heart to experience a constant state of adrenaline shock. The company was so successful that in 2012 it was acquired by Janssen, part of Johnson & Johnson.

*"I think that start-ups should be thought of less like companies and more like projects."*

**What else do you think would help to boost commercialization?**
I think that start-ups should be thought of less like companies and more like projects. You have an early-stage product, then you put it in the hands of an experienced project manager with all kinds of connections, who brings in the different partners. So we would form lean, virtual companies — investors don't want to finance Christmas parties and annual leave. But this sort of thing is harder to do in Germany than in places such as the United States or the United Kingdom because of the country's tough employment laws.

**How will BioM shape the future of the Munich cluster?**
In contrast to other cluster organizations, we always knew that we were a cluster-development organization rather than just a cluster-management organization. Development is the fun part. We try to identify the trends, and bring companies and research institutes together to work on them. In the past few years, we've convinced the companies in our region that personalized medicine is the treatment of the future. I admire what Genomics England has been doing with its 100,000 Genome Project, and I'm trying to convince politicians and those that have the money to do something similar in Germany.

We have the freedom to identify trends and bring them to industry and academic institutions. I think that's important, especially in a country where sometimes things take longer than elsewhere. We need advocates in Germany who recognize new developments early and can convince the politicians to support, and the scientists to work, in the field.

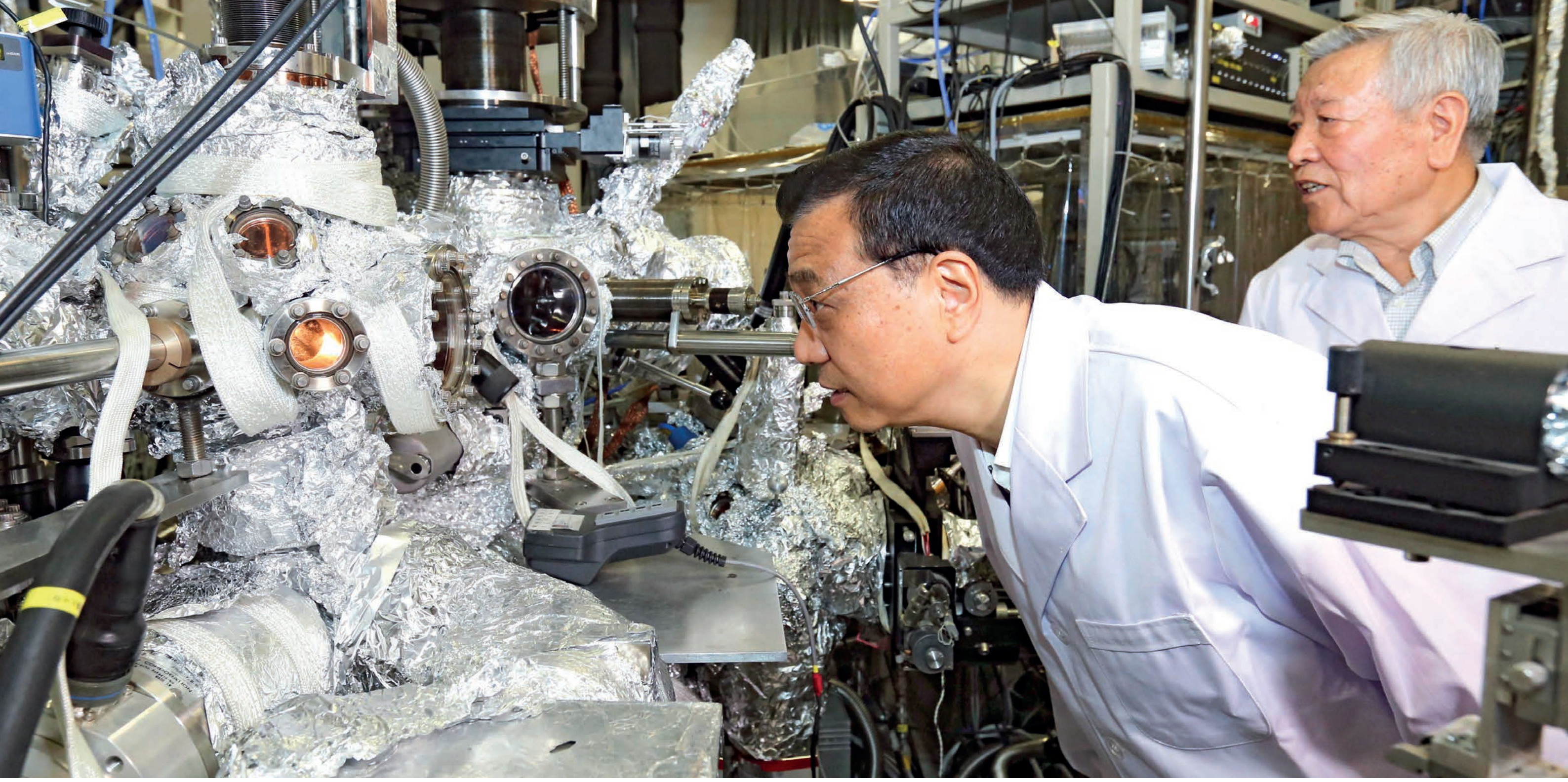
**Have you encountered political resistance in Germany?**
When scientists in Germany tried to start genetically engineering organisms in the late 1980s, they had to fight representatives of the Green Party. The party even opposed producing insulin using recombinant bacteria, which was commercialized in the United States in 1982. Then, in 1990, the German government passed a law to regulate genetic engineering, which satisfied the Greens, but created a huge bureaucratic burden for scientists. The next big challenge was agricultural biotechnology, or green biotechnology as we say in Germany. There, green biotech lost. The field is more or less dead. But the opponents are happy, at least — they no longer accuse the biotech industry of being one of the biggest dangers in Germany.

**Are you optimistic about the future?**
We have excellent science, but we do not have the investors that we need. We get companies started, but it's extremely difficult to raise the €10 million to €30 million needed to develop the technology. Scientists who could launch new businesses say, why should we, when we have no chance of being financed later? The entrepreneurial spirit is drying out. And that's what we have to change.

**What do you see as the missing ingredient?**
I think to have a successful biotech industry, we need local investors. If we provide incentives to those who have the money to invest in technology, then we would have a much better chance. The government has not solved this problem, and that's affecting not only biotechnology, but all areas that need venture capital. Whenever a company is successful, that financial success almost never stays in Germany because the funds are coming from elsewhere — either a non-German company buys the start-up, or the money comes from investors outside the country. So it's not German investors who are successful at the end of the day. ■

**INTERVIEW BY CHELSEA WALD**

**This interview has been edited for clarity and length.**



DING LIN/XINHUA PRESS/CORBIS

Chinese Premier Li Keqiang (L) visits the National Lab for Superconductivity at the Institute of Physics of the Chinese Academy of Sciences in Beijing.

CHINA

# Building an innovator

*When it comes to translating its own research into practical applications, China falls short. A forum in Shanghai put the spotlight on ambitious plans to accelerate the process.*

BY NICKY PHILLIPS

From a portable ultrasound machine that improves diagnosis in rural hospitals to unmanned military drones, China has excelled at adapting ideas and technologies for its large domestic market. "From an industrial research perspective, what's really important is to translate cutting-edge science into real applications," says Xiangli Chen, general manager of General Electric China Technology Center in Shanghai. "And that we're very good at."

Chen was speaking at the 2015 International Forum: From Research to Innovation and Entrepreneurship in Shanghai, which was co-sponsored by the Shanghai Association for Science and Technology (SAST), the Chinese Academy of Sciences Shanghai branch, the Shanghai newspaper the *Wenhui Daily* and Springer Nature (publisher of *Nature*). Scientists, policymakers and leaders of academia and industry gathered at the meeting to discuss how China can build a sustainable innovation ecosystem. Although the nation has mastered the art of tinkering and scaling up other countries' research and ideas, the forum discussed the elements China needs to transform its own scientific research into products, services and technologies.

China sees its future economic growth and social prosperity as being directly tied to how well it can innovate — and in particular, how well it can create truly new products driven by scientific research. The government's 2016 five-year plan lists this as its top priority.

Many countries have similar goals, but the scale of the Chinese government's investment and influence sets the country apart. In the past 15 years, China has more than doubled the percentage of its gross domestic product that it spends on research and development (R&D). National, provincial and local governments offer generous funding for almost anything related to innovation and entrepreneurship, and education reforms mean that schools are encouraged to foster the next generation of innovators. May Lee, the dean of the School of Entrepreneurship and Management at ShanghaiTech University, and a moderator at the forum, said that the country was well versed in scaling up programmes to capitalize on its large population. "Even if only 0.1% of the population are coming up with breakthroughs," Lee told *Nature*, "China can generate more than anyone else because the population is so big."

Despite China's ambition and investment, a 2015 report by the McKinsey Global Institute (E. Roth *et al. The China Effect on Global Innovation;* McKinsey & Company, 2015) found that the country's efforts were "yet to give China a lead in science-based innovation". The authors acknowledged that this kind of innovation often has a long lead time — it can, for instance, take years of investment to produce and commercialize new drugs or crop varieties. But they also found that under-investment in basic science, limited incentives for private R&D and regulatory bottlenecks that restricted market access to innovative products were all holding back the country's progress. Although some observers say that it is too early to assess whether China's top-down approach will provide a return on its substantial investment, others predict that only when the central government steps aside and allows market forces to prevail will science-based innovation truly flourish.

## HUMAN CAPITAL

"It's people who innovate," said Bernard Meyerson, IBM's chief innovation officer, during his keynote speech. "But it's a special type of person." Innovators must be experts in their field, but also able to communicate their ideas to their colleagues, grant providers, investors and consumers, he said.

China trains about 30,000 science and engineering PhD students each year, but they were educated in a system that valued published papers over the entrepreneurial skills that are valuable to industry. Companies

still complain of skill gaps among science, technology, engineering and maths graduates, according to the McKinsey report. But Lee says that's changing. A raft of education reforms to train students to be creative thinkers is being rolled out at the school and university levels. Changes to school curricula include increasing opportunities for students to be creative, curious and to learn by doing, she says.

It will take a decade or more for these initiatives to have a significant effect, and so the central government is encouraging thousands of Chinese scientists who are living abroad to return home to launch their own companies or labs. For now, the government is relying on these Chinese-born, Western-trained scientists and entrepreneurs to run labs and train PhD students and junior scientists. The '1000 Talents Plan' has far exceeded its eponymous goal. Since 2009, it has enticed more than 4,000 high-level scientists back to China with incentives such as high salaries. This has also meant paying to relocate many of the scientists' families, often from the United States, Europe, Japan and Singapore. Thomas Kenny, a mechanical engineer at Stanford University in California, says that returning scientists provide an "infusion of talent" that can have an immediate effect on the country's growing innovation ecosystem.

Several speakers at the forum, including Kenny, warned that China should avoid treating all innovators the same. Many universities in the United States expect scientists to be both great teachers and savvy entrepreneurs, but "often the skills and duties of an entrepreneur and teacher are conflicting", says Kenny, who has had both roles in his career. "Entrepreneurs and CEOs have to be ruthless; professors have to be mentors," he says. "It's dangerous to think those roles overlap much."

Cong Cao, a science-policy specialist at the University of Nottingham's campus in Ningbo, China, says that local faculty members face similar pressures. He is also concerned that the generous funding allowances offered to some academics may lead to conflicts of interest if they also try to set up their own start-up companies. "Maybe a professor uses his or her students or lab facilities for their business instead of using it for research at the university," Cao told *Nature*. He suggests that China should follow the Stanford model, which allows academics to take a break from academia when pursuing business opportunities, but return to their faculty positions later.

## BEHIND THE PACK

During a panel discussion on the part institutions can play in encouraging innovation, Jian Lu, vice-president of research at the City University of Hong Kong, said that good science would always find an application eventually, regardless of whether it was funded with specific applications in mind. But China has a paucity of basic research. An independent analysis by Nature Publishing Group called



Jian Lu of the City University of Hong Kong.

*Turning Point: Chinese Science in Transition* (see go.nature.com/ybsatt) found that China spends only 5% of its R&D expenditure on basic science, compared with 18% by the United States and 16% by the United Kingdom. In the report, four out of five high-level scientists interviewed agreed that China needed to invest more in basic research. "You've got to do basic science to have ideas that will eventually drive innovation in 10–20 years," said Chen, during a panel discussion at the Shanghai event.

Most Chinese R&D money goes to areas of research with more immediate commercial promise, and this has spurred a marked increase in patent applications over the past 15 years. Inventors in China applied for more than 1 million patents in 2015 — the fifth year in a row that the country has filed the highest number. Government subsidies, however, have contributed to the rise in applications, some of which represent low-quality patents that do not translate into commercial successes (J. Dang & K. Motohashi *China Economic Rev*. **35,** 137–155; 2015).

China's university technology-transfer offices still have a lot to learn about advising scientists and entrepreneurs on how to identify whether a patentable discovery has commercial value, says Ching Zhu, managing partner at venture-capital firm Frontline BioVentures in Shanghai. Although in 2013 China published 17% of the world's life-sciences papers, and in 2012 held 1 in 10 global life-sciences patents, the country launched only 2% of the world's new drugs that year. Chinese companies spend less on R&D as a percentage of their sales than their global competitors. One company that is generating knowledge is biotechnology giant BGI in Shenzhen, which employs more than 2,000 PhD graduates. It recently partnered with the Chinese Academy of Agricultural Sciences and the International Rice Research Institute to sequence 3,000 rice varieties, which will allow the fast-track development of disease-resistant plants.

## QUASHING IP THEFT

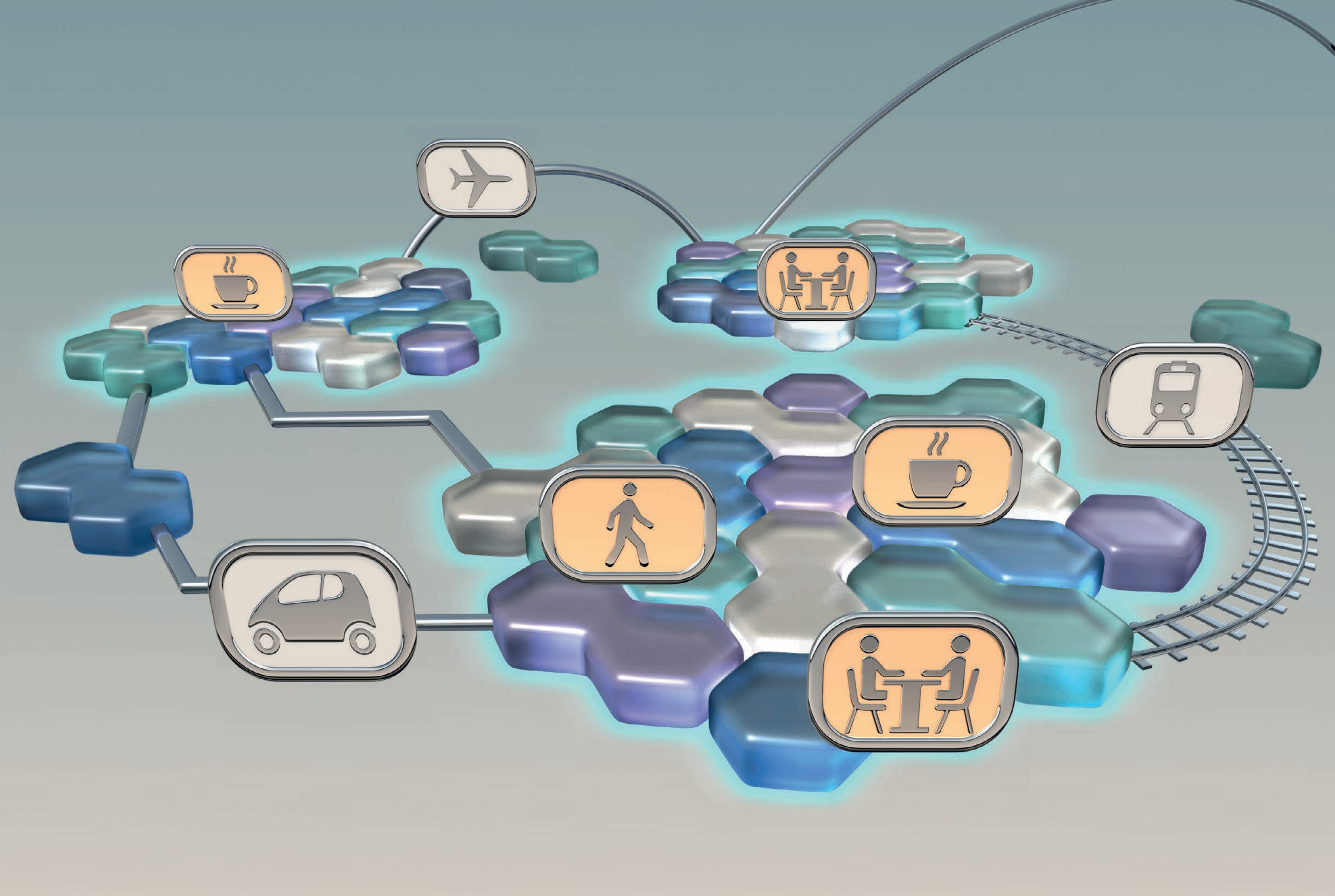
In the past, failure to effectively enforce intellectual property (IP) laws in China has deterred early-stage investors and venture capitalists from commercializing research discoveries. China is now trying to address this problem. "There is now a recognition and a serious effort on behalf of the government and industry to stamp out rampant IP theft," says Meyerson. Such policies will also help China's small, but growing, private venture-capital market to expand, says Zhu. China has plenty of capital, but lacks entrepreneurs who know how to connect scientists with potential investors — an essential component of a vibrant venture-capital industry — and this is keeping the market small. Investors would also like to see a more stable capital market. "In China, sometimes the market can be closed for months at a time," says Zhu. "There is a lot of government control." Zhu is confident that recent reforms such as reopening of the stock market to new listings and the Shanghai Stock Exchange's plan to introduce a market for small, innovative companies will increase investor confidence.

Despite China's efforts to adopt the policies that have led to the development of successful innovation cultures, such as those that helped to establish Silicon Valley, Cong says that in some instances, the Chinese government's involvement reduced the impetus for corporations to foster their own innovation pipelines driven by market needs. For instance, the central government's push to promote research commercialization has led to the construction of more than 100 high-tech science parks since the 1980s. But many of these hubs for science and industry collaboration have "homogenous development strategies", which have contributed to the oversupply of Chinese-made products such as solar- and wind-power technologies, says Cao. "Some things you can do top-down, but innovation you really need to have grow from the bottom, up." To promote innovation, he said, the role of government should be to create the right environment, with strong IP laws and funding for basic research.

But George Yip, a research specialist at China Europe International Business School in Shanghai, says that there are some obvious benefits to the central government's involvement: specifically, the size of the 88-million- member Communist Party. Through such a vast organization, the government can vigorously pursue its innovation goals at every level of Chinese society. "It's not just that China is top-down, it's that it reaches everywhere," said Yip.

Given China's current economic development, Lee also feels that the central government's approach is the right one. But she is not sure how long for, or how much it should shift to a system driven by entrepreneurs and the needs of the market. "Some amount of that must happen. But it could be very China specific and utterly different from what we've seen elsewhere." ■

**Nicky Phillips** *is a senior editor with Nature Publishing Group in Sydney, Australia.*



JOHN HARWOOD

COLLABORATION

# The geography of discovery

*Despite the ubiquity of the Internet, innovation still happens mainly in hubs, where face-to-face contact matters more than ever.*

BY EMILY SOHN

Damascus had steel. In Venice, it was glass. Switzerland made watches. For hundreds of years, regions developed specialities that often arose from access to a natural resource, but then intensified as people moved to the regions to be among the expertise.

The Internet was supposed to change all that. Around-the-clock connectivity that allowed researchers and entrepreneurs to collaborate from anywhere at any time meant that distance would no longer be an issue, predicted popular economic theory of the early 2000s. A decade later, it hasn't panned out that way.

Clusters of related and interconnected companies are stronger and growing more quickly than ever, innovation experts say — a trend that seems to be, paradoxically, fuelled by the Internet. Innovators and PhD students are now clumped together in fewer places, often in big cities. And collaborations are more likely to happen between researchers who live, or have lived, close to each other. "Lots of people want to write this story that clusters used to matter in the past, but they don't matter anymore," says Scott Stern, an economist of innovation and entrepreneurship at the Massachusetts Institute of Technology (MIT) in Cambridge. "For advanced economies and advanced science, location still seems to play a tremendously important role."

The classic example is Silicon Valley, "the mother of all clusters", says Martin Kenney, a geographer at the University of California, Davis. The region continues to attract technology workers and entrepreneurs at higher rates than other cities. But as data accumulate that show the importance of geography for spurring economic growth and signs of innovation (such as patent filings and scientific publications), economists have also uncovered examples of how distance, both physical and cultural, inspires discovery too.

Questions about how, when and why location matters are fuelling an active area of inquiry, with plenty at stake. And policymakers who want to stimulate economic development by attracting talent, boosting innovation and encouraging discovery are watching closely.

## PLACE MATTERS

The first economic treatise on the cluster phenomenon emerged in the late nineteenth century, when British economist Alfred Marshall described how concentrations of related businesses could be beneficial for the regions that host them. More recently, specialists such as economic geographers have taken a data-driven approach to try to understand the value of clusters, which form around all sorts of industries, from medical technology in Minnesota to mechanical engineering in Germany.

Overall, Stern says, dozens of studies show that clusters are good for both economies and for the generation of new ideas.

On the economic side, regions that host clusters have more jobs with salaries that grow more quickly compared with regions that don't host clusters — and not just within the speciality at the heart of the region. An analysis of data from 9-million workers across the United States found that every new high-tech and innovation job leads to the creation of 5 other jobs in the region, including lawyers, baristas, teachers and therapists. In places with lots of high-tech jobs, the result is many other jobs[1].

Regions with strong clusters are also more resilient in tough times. Stern and his colleagues found that US industries grew more during the 2007–09 recession if they belonged to an established cluster[2]. Within a strong cluster of medical-device manufacturers around Salt Lake City, Utah, for example, recession years saw annual employment grow by 5%, compared with an average decline of 14% across the United States and a drop of 31% around Madison, Wisconsin, which had few complementary industries. "During downturns," Stern says, "the financial guillotines hit hardest in regions where clusters are weakest."

New companies are more likely to form and start-ups are more likely to survive within clusters. When it comes to research advancements in science and technology, firms are more likely to file patents compared with more isolated companies. Writing in *Science*, Stern and Jorge Guzman, also at MIT, visualized this in a new way[3]. By focusing not on how many entrepreneurs there are in California, but on how likely those entrepreneurs are to be successful, they mapped where economic growth is likely to be the greatest — with a hot spot over Silicon Valley.

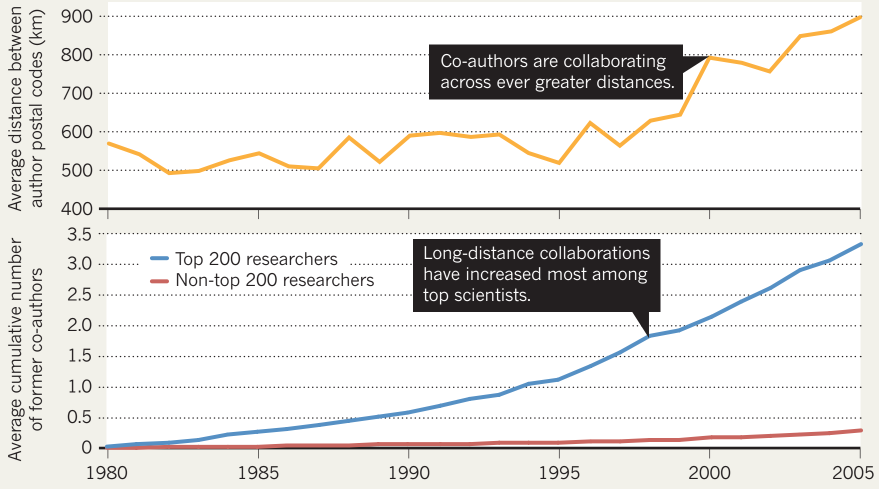
*"It's very critical to people's careers that they spend some time in clusters to connect."*

An analysis of the scientific literature also shows how success occurs in clusters. Patent filings and academic papers are more likely to cite other patents or publications that were produced nearby. When researchers looked at the citations of about 9,500 elite (frequently cited or highly funded, for example) life scientists, they found that when scholars moved to a new location, their work was cited less in patents by researchers in the place of departure, but cited more in articles by researchers living near the new destination, and face time seemed more important in industry than in academia[4]. "All indications are that proximity matters," says economist Paul Romer at New York University. "And it's possible it matters more now than it did in the past."

Serendipitous interactions with other researchers who might influence their work may be one reason why proximity remains so important, despite the ubiquity of the Internet,

## RESEARCH AT A DISTANCE

With the rise of the Internet (top), long-distance collaborations have increased. But distant relationships are most successful when authors have worked together before, especially for top-ranked researchers (bottom).




SOURCE: REF. 6

says Ajay Agrawal, a visiting economist at Stanford University in California. Agrawal studies collaborations. These have become increasingly important over the past few decades, as shown by the steady increase in the number of authors on papers — a trend that spans disciplines, from engineering to the life sciences.

When Agrawal analysed the geography of these ballooning federations of researchers, he found surprising clues about the foundations needed to encourage new discoveries, particularly in the context of the Internet age. Along with Avi Goldfarb at the University of Toronto, Canada, Agrawal found that between 1981 and 1991 (when many researchers began to use an early form of the Internet called Bitnet) the rate of collaborations between all institutions whose researchers published in top electrical engineering journals increased — but that collaborations were several times more likely to happen if researchers from each institution lived in the same US city[5]. That pattern persisted, he says, even as Internet use became universal.

And when scientists from different institutions do collaborate, there is a high chance that they once shared a physical space. For example, Agrawal and his colleagues found that typical collaborations among authors of articles in evolutionary-biology journals included teams of professors with former graduate students and postdocs who have since moved away[6]. (see 'Research at a distance'). "The lesson we learned from that is that in scientific discovery, a lot depends on relationships," Agrawal says. "Science is a social process."

That is not to say that online networks have no role in building relationships. Agrawal also found that the Internet has accentuated and accelerated the productivity that comes from face-to-face interactions. Researchers might meet for lunch or bump into each other in a hallway and end up discussing ideas, for example. But once they're back in their offices, they can easily follow up with e-mails and file-sharing. That kind of interplay between online and in-person communication, "means that it's very critical to people's careers that they spend some time in clusters to connect", he says. "When those relationships are established, they can go anywhere in the world." (see 'The power of face time').

And yet, the ease by which in-person meetings can be arranged still matters for long-distance collaborations. Christian Catalini, who studies the economics of innovation, entrepreneurship, and scientific productivity at MIT, and his colleagues found that the introduction of flights by low-cost US airline Southwest Airlines led to a 50% increase in paper collaborations by scientists at universities linked by those flights (see go.nature.com/fivxsr).

In a related analysis, MIT financial economist Xavier Giroud found that when airlines introduced new routes that decreased travel time from a company's headquarters to its remote plants, companies invested 8% more in those plants, and productivity increased[7]. And last year, Giroud and his colleagues found that direct flights make venture capitalists more likely to interact with their portfolio companies. The introduction of faster airline routes led to an increase of around 3% in the number of patents produced by the portfolio company and an increase of nearly 6% in the number of citations each patent received[8].

In one of the most illuminating and creative studies to illustrate the value of proximity, Catalini took advantage of a long-term clean-up project to remove asbestos from the largest medical and scientific complex in France — Paris's Pierre-and-Marie-Curie University. During the 15-year clean-up, which started in 1997, the university staged 5 major waves of relocations that involved moving lab groups around, essentially at random. Researchers had

## COMMUNICATION

### *The power of face time*

The buzz that drives people to move to industry clusters such as Silicon Valley feeds off the idea that 'being there' matters and that face time breeds trust. So, why can't we bond as well over Twitter, Facebook or Skype?

Brain chemistry might play a part, says Susan Pinker, a developmental psychologist based in Montreal and author of *The Village Effect: Why Face-to-Face Contact Matters* (Atlantic Books, 2014). Pinker makes the case for the widespread value of proximity; amid all the eye contact, handshaking and non-verbal communication that happens during meetings, our brains release a cascade of powerful neurotransmitters such as oxytocin and dopamine.

These feel-good messengers help us to lower our defences and become better able to assess the intentions and emotions of others, allowing us to build social cohesion and trust. Cooperation and teamwork then ensue. Only in person, she adds, can we read subtle social cues that reveal the trustworthiness of others. "The critical element has nothing to do with what workers are saying, and it can't be communicated via text or email," Pinker writes. "You have to be there."

Also hard-wired into our brains are finely tuned mirror neurons that cause people to mimic and subsequently like each other, adds Ben Waber, a people analytics researcher and visiting scientist at Massachusetts Institute of Technology in Cambridge. And face time seems to enhance the chances of those interactions occurring. When he and colleagues restructured coffee breaks at a bank's call centre to be synchronized instead of staggered, they discovered that employees became more productive, less stressed and more satisfied at work because they were able to support each other[13]. Workers whose breaks were staggered and who communicated mainly over e-mail did not show the same improvements.

"There are elements of face-to-face interaction that no technology has been able to effectively mimic," says Waber. "There's just a lack of richness of the media such that it doesn't really allow us to have these kinds of relationships." **E.S.**

no control over where they ended up, and they were given little notice of the moves.

Catalini found that these relocations had a major effect on both collaborations and publications. After getting shuffled around, lab groups became 3.5 times more likely to collaborate with their new neighbours and 5 times more likely to publish in a journal that was new for at least one of the collaborators[9]. The collaborations were also more likely to publish in higher quality journals. And although the study couldn't explain exactly why this happens, Catalini suspects that proximity simply reduces the "opportunity cost" of meeting up, in turn increasing the potential for more interactions and conversations that might lead to new ideas for research. "You could imagine that once people get co-located, grabbing coffee or having a conversation is less costly," he says. "They're more likely to engage in this exploratory behaviour."

And even for papers that did not involve neighbouring groups, the influence of these new relationships is evident. Catalini analysed author keywords of around 39,000 papers published by the complex's labs over 30 years. Compared with papers published before relocations, he found a 44% increase in keyword overlap among papers published 5 years or more after labs were placed near each other.

These collaborations can have big impacts. For instance, articles with four or fewer authors that were published by Harvard researchers in the same building were cited 45% more than were papers by authors working in different buildings[10]. Stories abound of research collaborations that formed because two people happened to connect and hit it off. Molecular biologist Herbert Boyer and geneticist Stanley Cohen teamed up to create the first recombinant organism after they ended up talking over a late-night snack at a deli in Hawaii, where they first met at a conference. Robert Solow, an economics Nobel laureate who was one of the researchers shuffled around in Catalini's Paris study, said at the time, "The truth is, it may have changed my whole life." His office relocation led to a friendship with fellow laureate Paul Samuelson, a relationship that steered him away from statistics and towards straight economics. "The location of that office and the fact that we liked each other so much had a major influence on the direction my career took."

### DISTANCE PERKS

Although clusters persist as important drivers of economic growth and innovation, the Internet means that distance has an important role in scientific discovery, too. After researchers move, many collaborations between separated colleagues drop off. But the most worthwhile relationships continue, Catalini says. And web-based programs, including e-mail, Slack and Twitter, are essential to making those relationships work.

Clusters with relatively few companies may benefit in particular from connections beyond their borders, says Rune Dahl Fitjar, an economic geographer at the University of Stavanger in Norway. To test the idea that proximity accelerates innovation, Fitjar and Andrés Rodríguez-Pose of the London School of Economics surveyed chief executives of more than 500 Norwegian companies in 2013 (ref. 11). The executives answered questions about their firms' levels of innovation, including the kinds of collaborations that they engaged in and the numbers of new products that they had introduced.

Fitjar and Rodríguez-Pose's findings were unexpected. For this group of Norwegian companies, which included hotels and manufacturing, construction and communication firms, regionally clustered collaborations failed to spark innovation. Instead, innovation was much more likely when firms collaborated with companies in other countries. The chief executives also said that meetings with colleagues are usually purposeful and planned, not random and accidental, suggesting that innovation is often as deliberate as it is serendipitous.

Physical geography isn't the only type of distance worth considering, Fitjar adds. He and colleagues asked Norwegian firms about their main partners in innovation, and uncovered what he calls a Goldilocks principle[12]. The most successful collaborations occurred between partners that were neither too alike nor too different in their values, attitudes, social structures and ways of thinking.

"What we see is that firms and innovators depend on these long-distance connections to innovate," Fitjar says "It's kind of a new story that hasn't been told in the literature before." ■

**Emily Sohn** *is a freelance journalist based in Minneapolis, Minnesota.*


1. Moretti, E. *The New Geography of Jobs* (Houghton Mifflin Harcourt, 2012).
2. Delgado, M., Porter, M. E. & Stern S. *Clusters and the Great Recession* (2015); available at http://go.nature.com/x7erzf
3. Guzman, J. & Stern, S. *Science* **347,** 606–609 (2015).
4. Azoulay, P., Graff Zivin, J. S. & Sampat, B. N. in *The Rate and Direction of Inventive Activity Revisited* (eds Lerner, J. & Stern, S.) 107–155 (NBER, 2012).
5. Agrawal, A. & Goldfarb. A. *Am. Econ. Rev.* **98,** 1578–1590 (2008).
6. Agrawal, A., McHale, J. & Oettl, A. in *The Changing Frontier: Rethinking Science and Innovation Policy* 75–102 (NBER, 2014).
7. Giroud, X. Q. *J. Econ.* **128,** 861–915 (2012).
8. Bernstein, S., Giroud, X. & Townsend, R. R. *J. Finance* http://dx.doi.org/10.1111/jofi.12370 (2015).
9. Catalini, C. Rotman School of Management Working Paper No. 2126890 http://dx.doi.org/10.2139/ssrn.2126890 (2015).
10. Lee, K., Brownstein, J. S., Mills, R. G. & Kohane, I. S. *PLoS ONE* **5,** e14279 (2010).
11. Fitjar, R. D. & Rodríguez-Pose, A. CEPR Discussion Paper No. DP11067 (2016); available at http://go.nature.com/qetfjc
12. Fitjar, R. D., Huber, F. & Rodríguez-Pose, A. *Not Too Close, Not Too Far* (2015); available at http://go.nature.com/gcnrxh
13. Waber, B. N., Olguín Olguín, D., Kim, T. & Pentland, A. in *Proc. 30th International Sunbelt Social Network Conference* (INSNA, 2010).

PIUS UTOMI EKPEI/AFP/GETTY IMAGES

Venture philanthropist Bill Gates looks on as a health worker vaccinates a child in Ghana.

FUNDING

# Donor drugs

*For the past decade, venture philanthropists have been working to propel promising therapies and vaccines into the clinic, with some success.*

BY CASSANDRA WILLYARD

In December 2015, Facebook founder Mark Zuckerberg and his wife Priscilla Chan made a stunning announcement. In a letter to their newborn daughter, the couple pledged to give away 99% of their Facebook stock — about US$45 billion. Part of that money, they said, will go towards finding therapies for five global killers: heart disease, cancer, stroke and neurodegenerative and infectious diseases. “Curing disease will take time,” Zuckerberg wrote. “Over short periods of five or ten years, it may not seem like we’re making much of a difference. But over the long term, seeds planted now will grow, and one day, you or your children will see what we can only imagine: a world without suffering from disease.”

But Zuckerberg and Chan aren’t inclined to simply write a cheque. They are part of a cadre of philanthropists taking a more hands-on approach. These venture philanthropists hope to leverage their business savvy to shepherd new therapies to market — fast. “They want to roll up their sleeves and understand how their dollars are being used to address unmet needs, to overcome research roadblocks and to take advantage of promising new discoveries,” says Melissa Stevens, executive director of the Center for Strategic Philanthropy at the Milken Institute in Washington DC. Max Wallace, chief executive of Accelerate Brain Cancer Cure, or $ABC^2$, in Washington DC, puts it more bluntly: “These type of new rich don’t want to look like fools. They don’t want their money to be wasted.”

Microsoft co-founder Bill Gates has become the poster child for venture philanthropy. Since 2000, the Bill & Melinda Gates Foundation has poured more than $20 billion into global health. But many of the basic tenets of the model arose more than a century ago. “When philanthropy was developing in America there was this idea that foundations had this great capacity, because they weren’t the government, to solve social issues and be really innovative and take big risks,” says Alexandra Graddy-Reed, who studies non-profit organizations and their policies at the University of Southern California in Los Angeles. “Carnegie and Rockefeller had many of these principles when they were giving a hundred years ago.”

In the past decade venture philanthropy has experienced a resurgence, with many foundations focused on new therapies. But the attributes that make this type of funding so effective can also stir up controversy or raise ethical questions. Philanthropic foundations are not accountable to the public, and some critics question whether wealthy benefactors have too much sway in medicine.

### TAKING RISKS

The US National Institutes of Health invests about $32 billion in biomedical research and development each year, much of which goes towards basic research. If a new therapy looks promising, “the expectation was that for-profit

venture capitalist companies would come in and help those academic researchers spin out successful cures or drugs," says Graddy-Reed, and then big pharmaceutical companies would take over. But in recent years, the system has broken down. According to a report by US trade association the Biotechnology Industry Organization, venture funding of private drug-development companies peaked in 2007 at $5 billion. Then the financial crisis hit and funding fell by nearly half, to $2.8 billion in 2010. Investments have begun to recover, but they were still below pre-crisis levels in 2014. Although 2015 was a banner year for drug and biotech companies seeking venture capital, industry experts point out that early-stage research is still underfunded. Increasingly, venture philanthropy is stepping in to fill the gap. "It has emerged as the industry's new high-risk capital," Stevens says.

***"These type of new rich don't want to look like fools. They don't want their money to be wasted."***

In many ways, philanthropic funding is well suited to the task. "We can take risks that neither governments nor the private sector can afford to take. We don't have the same pressures for monetary return," says Penny Heaton, director of vaccine development at the Bill & Melinda Gates Foundation. "Our metrics are all about saving lives." This appetite for risk allows foundations to fund early-stage drug development, and even support unorthodox approaches. "They can really feed exploration in scientific areas where others might not be willing to go because it's so new or so innovative," Stevens says. For example, the Stanley Medical Research Institute — founded in 1989 by Ted and Vada Stanley, whose son was diagnosed with bipolar disorder — supports research that investigates infectious agents such as the parasite *Toxoplasma gondii* as possible cause of schizophrenia. "When we started our research on infectious agents 25 years ago, it would have been impossible to get government funding," says psychiatrist E. Fuller Torrey, associate director for research at the institute.

But moving drugs through the pipeline takes more than funding. "If money were the solution, I think this problem would have been tackled long ago," says Jonathan Stamler, director of the Harrington Discovery Institute in Cleveland, Ohio.

Stamler was a cardiovascular researcher at Duke University in Durham, North Carolina, when the financial crisis hit, and he watched with alarm as funding for drug development dried up. "It became increasingly difficult to find a way to move discovery forward," he says. So Stamler came up with an approach to fund early-stage innovators — an organization that would behave as both a non-profit institute and a for-profit company.

After relocating to University Hospitals Case Medical Center in Cleveland in 2010, Stamler paired up with Baiju Shah — who had extensive experience of launching biomedical companies — and together they took Stamler's idea to local philanthropist Ronald Harrington. Harrington and his family had already donated money to support cardiovascular research after he had a quadruple bypass in 2000. But Stamler and Shah pitched a way for the Harringtons to have an even greater impact on medicine: the non-profit institute would provide researchers with funding and much-needed industry expertise, and the for-profit accelerator would develop the most promising discoveries and hand them off to pharmaceutical firms to carry forward.

"They came at us four times," Harrington recalls. Eventually, the family agreed, and the Harrington Project for Discovery and Development was born. The Harringtons donated $50 million to kick-start the non-profit arm, the Harrington Discovery Institute, and drummed up another $100 million in support from other donors. The Harringtons also invested an undisclosed, but much smaller, amount in the for-profit arm, BioMotiv. Harrington may have been sceptical at the outset, but he has since become an champion. "This opens up collaboration like no other model," he says.

The project is just four years old, but already BioMotiv has brokered deals with several major pharmaceutical companies. Goutham Narla, a medical geneticist at Case Western Reserve University in Cleveland, thinks that his discoveries would have languished if he hadn't been selected as a 2012 Harrington Distinguished Scholar. "We just don't have the depth of pharmaceutical expertise in academia to do what I'd call true drug development," he says. The Harrington Discovery Institute helped Narla to develop an anticancer therapy, and now his company, Dual Therapeutics, is part of BioMotiv. "We have weekly calls with people who have, collectively, 80-plus years of experience in pharma," he says, and access to that experience is paying off. In January, BioMotiv announced that Dual Therapeutics would partner with drug giant Bristol-Myers Squibb. "The goal is to hopefully do clinical trials next year," Narla says.

**Priscilla Chan and her husband Mark Zuckerberg.**

PEOPLEPICTURE/PETRA PREUSS/REX/SHUTTERSTOCK

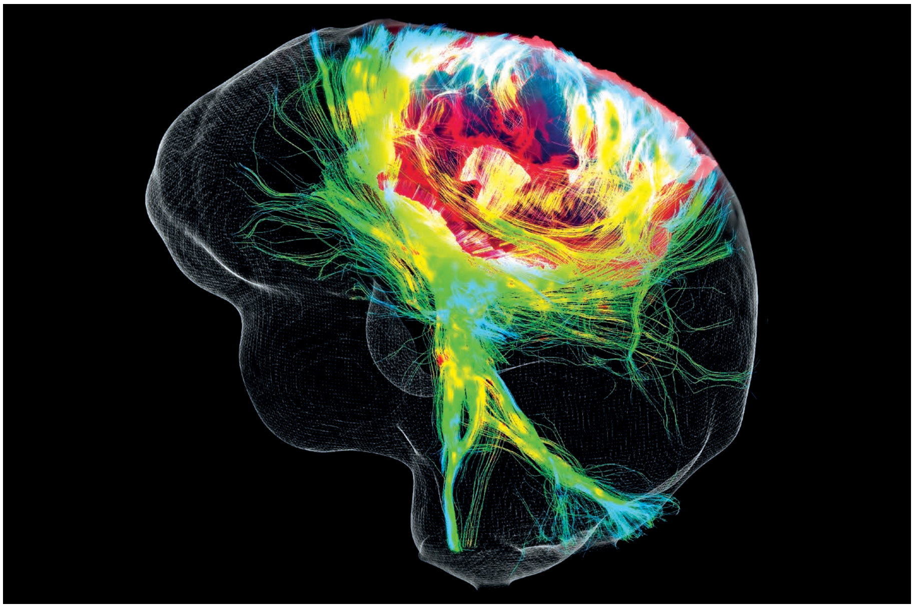
Many foundations have their roots in personal tragedy. As Stevens and her colleagues at the Center for Strategic Philanthropy like to say, "You don't go to medical philanthropy — medical philanthropy comes to you." For the Case family, tragedy struck in 2001, when 43-year-old investment banker Daniel Case was diagnosed with an aggressive type of brain tumour called a glioblastoma. "2001 doesn't seem like that long ago. But in brain-cancer terms, it's kind of the dark ages," Wallace says. Finding out that there were not any drugs available, Case enlisted the help of his brother Steve, the co-founder of digital media company AOL. Together with their families, the brothers founded ABC$^2$.

Like Stamler, Wallace doesn't see money as the main barrier to drug development. ABC$^2$, a 5-person foundation, has handed out only about $22 million in grants since 2001, and Wallace says that these days the foundation spends just $2–3 million a year. That may not be enough to fund clinical trials, but the money helps to bring people together. "Our role has often been to be a bio-yenta. Let's make some marriages," he says. One particularly fruitful marriage began at the ABC$^2$ 2012 Annual Scientific Meeting in Sausalito, California. Wallace and his colleagues struck up a conversation with William Sellers, global head of oncology at Novartis Institutes for BioMedical Research, headquartered in Cambridge, Massachusetts. Novartis had been looking at combination therapies for cancer, but it wasn't developing any drugs for brain tumours. When Wallace asked why, Sellers explained that they needed tumour tissue.

Brain-tumour samples can be tricky to extract, but Wallace knew that neurosurgeons at the Henry Ford Hospital in Detroit, Michigan, had a reputation for having 'magic hands'. So Wallace asked Tom Mikkelsen, co-director of the Hermelin Brain Tumor Center at Henry Ford, to join the discussion. Within days, Mikkelsen had samples ready for Novartis. These allowed the company to generate mice with human brain tumours that they could use to screen Novartis's compound library for therapies. "We're really small, but we're trying to cast a big shadow,"

Wallace says. "We've made $20 million worth of grants, but we have backed research that's led to 14 therapies being in the clinic."

SHERBROOKE CONNECTIVITY IMAGING LAB/SPL



Being diagnosed with a glioblastoma (pictured) prompted banker Daniel Case to set up a foundation.

## CAUTIONARY TALE

Although additional funding for biomedical research may seem like a winning proposition, the attributes that make philanthropic funding so powerful can also become stumbling blocks. Foundations can take more risks because they aren't accountable to the public. But this lack of accountability can also be a cause for concern. "We've given all these organizations, these individuals, a huge tax reprieve," Graddy-Reed says. "But the public has no say in how the foundations spend their money."

In some cases, philanthropic organizations are so large that they can drive the research agenda in a given area by themselves. The Gates Foundation, for example, has given away more than $36 billion since its inception. Around half of that has gone to global health, making the foundation the largest private supporter in that arena. "Their budget dwarfs the budget of small countries," says Gregg Gonsalves, a researcher at Yale Law School in New Haven, Connecticut. In 2014, the Gates Foundation supplied $2.9 billion (or 8%) of the $35.9 billion that high-income countries provided to support global health. In some cases, the foundation's impact is even larger. The same year, the foundation gave 13.9% of the total funding for maternal, newborn and child health, and 12.6% of the total funding for tuberculosis.

"Because of their size, they have this huge ability to influence what it is we're trying to do as a society," says Graddy-Reed. Whether that influence is a boon or a burden is a matter of debate. In a 2008 memo obtained by the *New York Times*, the then-director of the World Health Organization's malaria programme Arata Kochi, wrote that the foundation's tendency to push its favourite research "could have implicitly dangerous consequences on the policymaking process in world health". Kochi is one the few outspoken critics of the foundation. In certain fields, nearly everybody has some involvement with the Gates Foundation, Gonsalves says, and "nobody is going to want to bite the hand that feeds them."

Gonsalves acknowledges that the Gates Foundation has done a lot of admirable work, but he worries about the influence it could be having at a global level. The foundation is one of the largest funders of the World Health Organization. In 2014–15, it gave the organization $423 million — less than the United States, the agency's biggest donor, but more than the United Kingdom donated. That money is earmarked for particular projects. For example, nearly 70% of the Gates Foundation's 2014–15 contribution went to polio eradication. Ultimately, philanthropists have their own viewpoints and priorities, Gonsalves points out, and those drive the research agenda of their foundations.

Venture-philanthropy funding often comes with strings that can make commercialization more difficult, according to Kelly Sexton, director of the Office of Technology Transfer at North Carolina State University in Raleigh. Foundations might want royalties that are so high that the university is left empty handed. Or they may stipulate that the research be allowed to be licensed to multiple parties, which means that "when we go to find a licensee, we can't offer an exclusive licence", Sexton says. That can make it next to impossible to find companies willing to take up the technology.

When a foundation benefits financially from a drug that it helped to develop, the ethics can be murky. Since 2000, the Cystic Fibrosis Foundation has poured $150 million into Vertex Pharmaceuticals and another company that Vertex acquired to develop new drugs for the disease. In return, the foundation negotiated to keep some royalties. The investment paid off in 2012, when the US Food and Drug Administration approved Kalydeco (ivacaftor) — the first drug to treat the underlying cause of some forms of cystic fibrosis. "Funds from any royalties we receive are reinvested into further research and drug development and advance our mission to find a cure," the president and chief executive at the time Robert Beall wrote in 2014, when the royalty rights were sold for $3.3 billion.

***"Because of their size, they have this huge ability to influence what it is we're trying to do as a society."***

Some see the Kalydeco example as a success story of how non-profits can work with industry to bring much-needed drugs to market. But others have criticized the foundation for failing to negotiate a better price for the drug, which costs about $300,000 per person annually. Lisa Schwartz, a medical-communication researcher at Dartmouth Institute for Health, Policy and Clinical Practice in Lebanon, New Hampshire, notes that developing effective drugs is only part of the equation. "Isn't there some responsibility to provide access?" she asks.

Schwartz also says that having a stake in the sale of a drug intended for the patient group you're trying to serve creates a conflict of interest. "If you personally benefit every time that drug is prescribed, then the question is, will you fairly represent that drug?" she says. "Can you be the honest broker?"

The Cystic Fibrosis Foundation deal has prompted other organizations to ask whether they should adopt a similar model. "I am certain that after the announcement of the sale, every medical-research foundation in the US had a call with their board," says Stevens. But despite the potential windfall, some have decided to forgo the profits. "They don't want to run the risk of being seen as a non-neutral party," she says.

Zuckerberg and Chan's philanthropic plan is even more controversial. Rather than launching a foundation, the couple has set up a limited-liability company whose mission is "advancing human potential and promoting equality," according to the couple. Unlike a foundation, a limited-liability company can freely invest in for-profit organizations without disclosing those investments, make political contributions and lobby governments. By doing so, Zuckerberg and Chan sidestep the restrictions that govern charitable foundations. Zuckerberg argues that this will give the couple "the flexibility to give to the organizations that will do the best work — regardless of how they're structured", he wrote in December. No one can fault the pair for wanting to cure disease, but it remains to be seen whether this flexibility will lead to faster cures or just more controversy. ■

**Cassandra Willyard** *is a freelance science writer in Madison, Wisconsin.*

MORITZ HAGER/WORLD ECONOMIC FORUM


# Q&A Helga Nowotny

# Embrace uncertainty

*Austrian social scientist Helga Nowotny was president of the European Research Council between 2010 and 2013. Now a professor emerita of ETH Zurich and author of* The Cunning of Uncertainty *(Polity, 2015), Nowotny discusses the growing pressure to capitalize on academic research, and how countries can get it right in the absence of a universal recipe.*

**What are the factors that drive the push for more technology transfer and commercialization?**
Entrepreneurship has become infectious. Young people dream of setting up a company. The idea of bringing their scientific skills and knowledge to the market is gaining traction. I see two driving forces behind this. The first is that many more opportunities exist at the interface of science, technology and innovation today than there were 15 years ago. The second is the realization by the young that if they want to have jobs in the future, they must engage in creating them.

**How is this push changing how science is done?**
It is not so much about changing scientific fields, but about crossing fields. A new kind of practical interdisciplinarity is in the making. I saw it happen with the European Research Council (ERC) Proof of Concept scheme. These grants allow researchers who are already being funded by the ERC to explore the innovation potential of their research and to move towards its commercialization. The scheme awards up to €150,000 (US$169,500) per grant. Recent winning projects include super-hard fibres produced by bionic silkworms and artificial veins inspired by marine sponges. We also saw that it is not so much the ERC grantees who want to set up their own firms, but their talented PhD students and postdocs, and the scheme provides great opportunities for those people to do so. Young researchers understand that the boundaries between academic research and its practical uses are more porous than often thought. We need to provide the training to help them to make the leap to the other side.

**What are the pitfalls of commercializing research? How can they be avoided?**
I don't think there's one right way to go about promoting technology transfer, but there are common pitfalls. Timing is obviously important — one can be too early or come too late. Another factor is how to obtain financing between the initial phase, when it is easy to obtain money because the sums are small, and the later phase, when funders are rare and hesitant because the sums are larger. The barriers to scaling up have been highlighted as a problem. How can one move from having many small firms to having a few with the real capacity to grow? Young people must realize that neither their technological know-how nor their enthusiasm are sufficient. To have an idea is only the beginning. They also need knowledge of business models, modes of financing and what the market looks like.

**How can a small country, such as Austria, encourage research commercialization?**
Small countries often feature hidden champions — companies that do extremely well by operating at a global level in a technological niche. Doppelmayr/Garaventa is an Austrian example. It makes ski lifts in Austria and is now providing horizontal lifts for cable cars in cities around the world. The European Research Area Council Forum Austria was concerned that Austrian research and innovation systems were losing their dynamism, so it commissioned a study. The report (see go.nature.com/ugnlju), which it presented in November 2015, compared Austria's research, higher-education and innovation system with those of Denmark and Sweden.

**What did you conclude?**
First, there is no recipe for how to become an innovation leader. But the report helped us to see the interconnections in the research–education–innovation ecosystem more clearly. It recommended a more systemic evaluation of the effects that the present mix of policies generates. Better alignment of the well-intentioned, but often separate, efforts of the many players will be necessary if a small country is to succeed in a global world. Sweden and Denmark invest more in higher education than does Austria, and they do a better job at linking funding of higher-education institutions to the number of student places. Is the Austrian division between the general education offered by universities and the professional training offered by the *Fachhochschulen* (vocational universities) optimal?

***The Cunning of Uncertainty* calls for scientists to embrace uncertainty. But how can they do that when under pressure to seek profits?**
Politicians think in the short-term. They want to see predictable and almost immediate results with high economic impact. But fundamental research is an inherently uncertain process. It reaches out into the unknown, discovering what nobody thought existed or would be possible. Innovation is also an inherently uncertain process. It is important to see that uncertainty is the invisible ally of both fundamental research and innovation, and, if we embrace it, we have nothing to fear from it. Funders must make room for the different types of uncertainty and encourage scientists to capture the opportunities that they offer. Profit-seeking comes later, and it belongs to the market. ■

**INTERVIEW BY CHELSEA WALD**

**This interview has been edited for length and clarity. Published online 29 April 2016.**

ILLUSTRATION BY ELIOT WYATT

# Companies on campus

Housing industry labs in academic settings benefits all parties, say **Jana J. Watson-Capps** and **Thomas R. Cech**.

Pete Mariner works up the hall from his PhD adviser and one floor down from his postdoc adviser, but he does not work in academia. He is a senior scientist at Mosaic Biosciences, a start-up developing synthetic materials to help wounds heal faster, yet his labs are in the University of Colorado Boulder. They are part of the university's BioFrontiers Institute, an interdisciplinary effort to tackle complex biology and forge connections with companies.

Over the past three decades, academia and industry have been converging philosophically and physically[1]. Thirty-four years ago, the Bayh–Dole Act encouraged US academics to patent their discoveries, work with companies and become entrepreneurs[2]. Policies in Europe have moved in similar directions[3]. Companies increasingly partner with university scientists to enhance their research. In a 2007 survey of life-sciences faculty members from the 50 US universities that receive the most financial support from US National Institutes of Health, just over half of the respondents reported having some relationship with industry[4].

THE UNIVERSITY EXPERIMENT
A *Nature* special issue
nature.com/universities

Successful academia–industry partnerships require common interests, trust and good communication. For each of these, proximity helps.

Many universities have off-campus research parks, but some academic research facilities have gone a step further and brought small companies within their own walls. BioFrontiers (of which J.J.W.-C. is associate director, and T.R.C. is director) is one of the youngest experiments in 'co-location'. More are set to open soon (see 'Within the same walls'). When it is done well, all parties benefit.

## BUILDING BUDDIES

Various university offices connect faculty members, students and companies through technology transfer, industrial partnerships, student internships and mentoring. But these centralized resources do not allow for the spontaneous interactions that can arise from shared excitement about solving a problem. Co-location removes the physical separation and the intermediaries between researchers in academia and those in industry, and so allows serendipitous relationships to bloom.

Faculty members benefit from the influx of corporate expertise[5]. Researchers with industrial experience are often more knowledgeable about high-throughput technology and commercial applications than their academic counterparts. Our biomedical faculty members tell us that they value industry collaborations as a way to apply discoveries in ways that eventually benefit patients. Students gain real-world experience and opportunities to work at these companies as they expand. Young companies benefit from access to flexible lab space, core facilities, an invigorating research environment and an educated workforce.

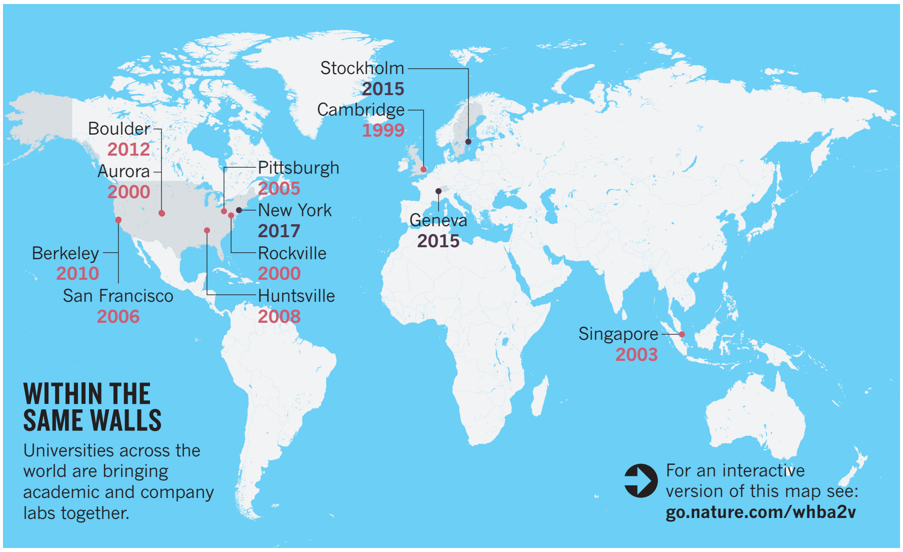
For example, when start-up Archer Dx, based in Boulder, began developing next-generation sequencing kits and software to research cancer treatments, it kept capital expenditures down by renting pre-built lab space at BioFrontiers and buying services from the university's genomics facility. When the company was purchased by a larger diagnostics and reagents company (Enzymatics, headquartered in Beverly, Massachusetts) and moved to a larger space off campus, it hired several former students.

Another example of co-location is the California Institute for Quantitative Biosciences (QB3). This supports two on-campus incubators for University of California spin-out companies, called 'biotech garages' in homage to the early Silicon Valley tech start-ups. One QB3 start-up is Caribou Biosciences, founded on ▶

COMMENT

genome-engineering technology from Jennifer Doudna's lab at the University of California, Berkeley. Following a now-familiar pattern, Caribou began operations in the Garage@Berkeley — steps from the Doudna lab — before moving into a larger space as the company grew.

HudsonAlpha Institute for Biotechnology, a non-profit organization in Huntsville, Alabama, brings together principal investigators, postdocs and some students alongside core facilities and independent companies that are developing new genomic technologies. ThermoFisher Scientific, a global biotech company based in Waltham, Massachusetts, bought one of the institute's start-ups in 2008, and retains its operations in Huntsville, citing the importance of proximity to researchers outside their own expertise.




**WITHIN THE SAME WALLS**
Universities across the world are bringing academic and company labs together.

### RULES OF ENGAGEMENT

Co-location has challenges. Universities are among the last places to prize research for the sake of pure discovery. All co-location leaders, business representatives, university administrators and development officers must help to implement the goals of the programme while protecting blue-sky research.

Ideally, co-location should be financed with funds that would not normally go to basic research, such as rent from tenant companies, philanthropic donations aimed at entrepreneurship and targeted grants. We have furnished several core facilities serving both academics and local companies using infrastructure grants from Colorado's Office of Economic Development and International Trade. HudsonAlpha was founded and largely funded by scientist-entrepreneurs Jim Hudson and Lonnie MacMillan, specifically to house academic faculty members alongside small companies. A*STAR (Agency for Science, Technology and Research) in Singapore is funded mainly by government programmes to boost commercial research and development.

Nonetheless, universities need to devote resources to addressing real and perceived conflicts of interest. This requires careful policies on intellectual property, use of university resources, faculty time and conflicts of interest. For example, students cannot be graded and employed part-time by the same person. On-campus companies should explicitly ensure participating students' ability to publish in a timely fashion, a practice already established for sponsored research agreements.

Companies predisposed to open science might be attracted to co-location. Accommodating these companies on campus demands flexibility and clarity. Just as universities need to be up-front about their goals and expectations, they also need mechanisms to remove participants who might be better off in more conventional settings. For example, we have offered leases on lab space as short as six months, which can be renewed. In the future, lease renewals at BioFrontiers might also depend on how companies interact with academic neighbours, for example through mentoring students.

Letting space to companies puts universities in the sometimes-awkward position of a landlord who needs to evaluate whether potential tenants can fulfil their rental payments and other obligations. Already, we have had a very young company leave a lab space after less than a month because anticipated seed funding did not come through.

### COOKIE HOUR

Customs and architecture should stimulate interactions. In the BioFrontiers building, academic and company researchers share a café and common spaces. Labs and offices are arranged so that people must pass through a main corridor to get from one to another, encouraging hallway conversations. Each week, a company or academic lab hosts a 'cookie hour' for anyone in the building. There are also whiteboards in hallways, where a spontaneous interaction can quickly turn into an idea sketch.

> ***"A university must view companies as partners in its research and education mission."***

Co-location will be most successful in academic settings that explicitly value entrepreneurship and translational research activities (for example, when recruiting faculty members or evaluating them for promotion and tenure), and where resources are available to foster community and to support a leadership team to oversee the programme. Emerging companies will be more likely to take advantage of co-location opportunities if there are grants and seed funds available to subsidize their rent, if core facilities are available and if research collaborations with the university are easy to set up.

Fundamentally, a university must view companies as partners in its research and education mission, not simply as an alternative revenue source.

### UNIVERSITY ECOSYSTEM

We believe that the daily interaction between education, research and enterprise resulting from co-location will connect universities to their communities and make them more relevant to students and parents paying tuition fees. Co-location sites will become magnets for entrepreneurial faculty members, postdocs and students, as well as for companies looking to hire new talent.

The intersection of academia and industry will become more natural as faculty members look for more ways to make their discoveries relevant, as students want more value for their degrees, and as companies want more input into developing their workforce. Industrial inhabitants will be part of the future university ecosystem. ■

**Jana J. Watson-Capps** *is associate director of the BioFrontiers Institute at the University of Colorado Boulder, USA.* **Thomas R. Cech** *is professor of chemistry and biochemistry at the University of Colorado Boulder and director of the BioFrontiers Institute.*
*e-mail: jana.watson-capps@colorado.edu*

1. Schachter, B. *Nature Biotechnol.* **30,** 944–952 (2012).
2. Grimaldi, R. *et al. Res. Policy* **40,** 1045–1057 (2011).
3. Perkmann, M. *et al. Res. Policy* **42,** 423–442 (2013).
4. Zinner, D. E. *et al. Health Aff.* **28,** 1814–1825 (2009).
5. D'Este, P. & Perkmann, M. *J. Technol. Transfer* **36,** 316–339 (2011).

# COMMENT

**ENERGY** Governments must keep backing the renewables boom **p.297**

**HOMININS** Was the hunt for water the key driver in human evolution? **p.303**

**VISUALIZATION** Exhibition celebrates history of scientific graphics **p.304**

**REPRODUCIBILITY** Hypothesis-driven projects risk data discounting **p.306**

ILLUSTRATION BY DAVID PARKINS

# Industry–funded academic inventions boost innovation

**Brian D. Wright** and colleagues present data challenging the assumption that corporate-funded academic research is less accessible and useful to others.

Governments have long encouraged university–industry collaboration, hoping to spur innovations that bring jobs, investment and life-enhancing products[1]. At the same time, shrinking government budgets for science have forced universities to look to other sources of funding. According to the US National Science Foundation, in 2012, industry supplied just over 5% (some US$3.2 billion) of US research universities' annual expenditure[2].

But the role of corporations in academic research is controversial. For example, when oil company BP announced in 2007 that it would pay $500 million to fund a decade of alternative-energy research by a consortium headed by the University of California, Berkeley, this prompted a backlash. Fearing that industry money would contaminate the public institution's research agenda, many students, staff and members of the community picketed the campus with a 2.5-metre Trojan horse. An earlier agreement between the department of plant and microbial biology at Berkeley and the Swiss pharmaceutical firm Novartis sparked similar opposition. At the 1999 graduation ceremony, about 100 students displayed the company's logo on their mortarboards, protesting that the department had been bought by corporate interests.

There are reasons to be cautious about corporate sponsorship of academic research[3]. The tobacco, food, pharmaceutical and other industries have been shown to manipulate research questions and public discourse for their own benefit and even to suppress unfavourable research[4]. And companies may shift university researchers towards narrow corporate interests. If the results of research are privately held, others cannot exploit them.

Conversely, some feel that overly restrictive university technology-transfer policies stifle productive deal-making between firms and academic researchers[5]. Some advocate that a university's intellectual property should ▶

be managed by an outside agency[6], or else handed over directly to researchers or to the companies funding their work[7].

Data to inform this debate are hard to come by. Individual universities may track patents and licences at their own institutions, but these data sets are generally small and confidential. The prevailing assumption is that corporate-sponsored inventions and the information associated with them are less accessible and less useful to others than inventions sponsored by the government or non-profit organizations.

Here we offer empirical evidence to the contrary. Our analysis suggests that corporate-sponsored research is surprisingly valuable for further innovation. Data collected over 20 years at nine campuses and three national laboratories administered by the University of California show that corporate-sponsored inventions are licensed and cited more often than federally sponsored ones.

Although results might differ at other academic institutions, these findings should allay concerns that corporate sponsorship turns leading universities into corporate vassals. Collecting and combining data from a larger sample of institutions could help to both explore what corporations hope to gain from funding academic work, and suggest how universities can best manage research sponsorships.

## TECHNOLOGY TRANSFER

Like most universities, the University of California requires faculty members and other researchers to disclose any invention that has commercial potential to one of its offices of technology transfer (OTTs), and to list funding sources for the project that led to it. Under these terms, an invention is anything that a researcher feels could be patented or is otherwise valuable as intellectual property: it might be a material, a method, or an animal or plant. The OTT then determines whether to pursue intellectual property protection on the university's behalf and negotiates contracts with potential licensees.

From 1990 to 2005, University of California faculty members, staff and students, and employees of the three associated national laboratories disclosed 12,516 inventions to their OTTs. Of these, nearly 1,500 were supported, at least in part, by corporate funds. Under strict terms of confidentiality, the central OTT provided us with data on these disclosures, and on related licensing activities, until the end of 2010. From 1990 to 2010 the University of California campuses accounted for up to 9% of total US academic research expenditure. Collectively, they obtained more issued patents than any other US academic institution. In lists compiled annually by the US Patent and Trademark Office, the multi-campus University of California system often had more than twice as many patents as the second-largest patent producer in academia (generally the Massachusetts Institute of Technology in Cambridge).

Of all inventions generated at the University of California, 20% are linked to at least one licence, and nearly 25% were eventually patented. Inventions with no sponsor information were the least likely to yield either licences (13%) or patents (17%). We believe that most of these inventions came either without extramural support or with federal support, which is such a common situation that inventors or technology-transfer agents may not note it explicitly. Corporate-sponsored inventions resulted in licences (29%) and patents (35%) more frequently than federally sponsored ones (22% and 26%, respectively). The rates are higher still for inventions with both types of sponsor; 36% were licensed and 43% patented (see 'Licensed and cited'). Results were similar across technical fields. More than two-thirds of classified technologies relate to biological, pharmaceutical and chemical advances, a distribution that is consistent with other leading research universities (for the complete results see Supplementary information; go.nature.com/o99eua).

Although corporate-sponsored inventions are more likely to be patented, that does not mean that corporate support makes inventions more patentable. Instead, corporations might select projects that are more likely to produce patentable inventions.

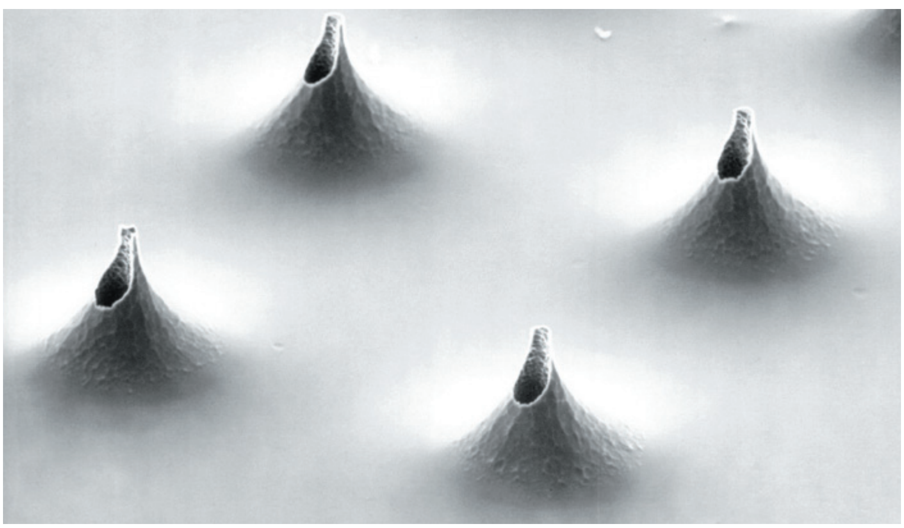
Corporations typically get priority to negotiate licences to the inventions they sponsor, and 86% of the licences to the sponsors are exclusive, meaning that the university agrees not to grant the same rights to multiple licensees. Of licensed inventions associated with some form of intellectual property, 78% were licensed exclusively, consistent with the share of 79% reported for licensing of patents funded by the National Institutes of Health[8].

Nevertheless, our analysis did not support our original assumptions that licences to industry-sponsored inventions would be likely to be exclusive, or that sponsors would snap up the lion's share of exclusive licences. First, the overall percentage of corporate-sponsored inventions licensed exclusively (74%) is not higher than for those with solely public funding (76%). Second, half of the exclusive licences for corporate-sponsored inventions seem to be to third parties (although we cannot be sure that we identified all the sponsor-controlled firms in the data). Apparently, even the inventions that sponsors leave on the table have substantial value, because these licensees usually bear significant costs of patenting, plus agreements to pay future royalties.

Another surprise is that corporate-sponsored inventions spur more 'knowledge spillovers', on average, than federally sponsored research, according to forward citation rates, the most widely used metric for patent quality and value. Forward citations show how many times one patent is cited in subsequent patents. Each corporate-sponsored invention generated, on average, 12.8 forward citations if licensed to a third party (more if licensed by the sponsor), compared with 5.6 for federally sponsored inventions. This runs counter to the expectation that corporate-sponsored inventions have narrow applications, and so create more private benefits but few benefits for others.

## USING UNIVERSITIES

This analysis does not address how corporate funds affect universities' research agendas, but it does dispute the idea that corporations tie up all sponsored inventions to restrict access. Instead, high patent citation rates for corporate-sponsored inventions suggest that firms are funding exploratory research. Work by sociologist James Evans



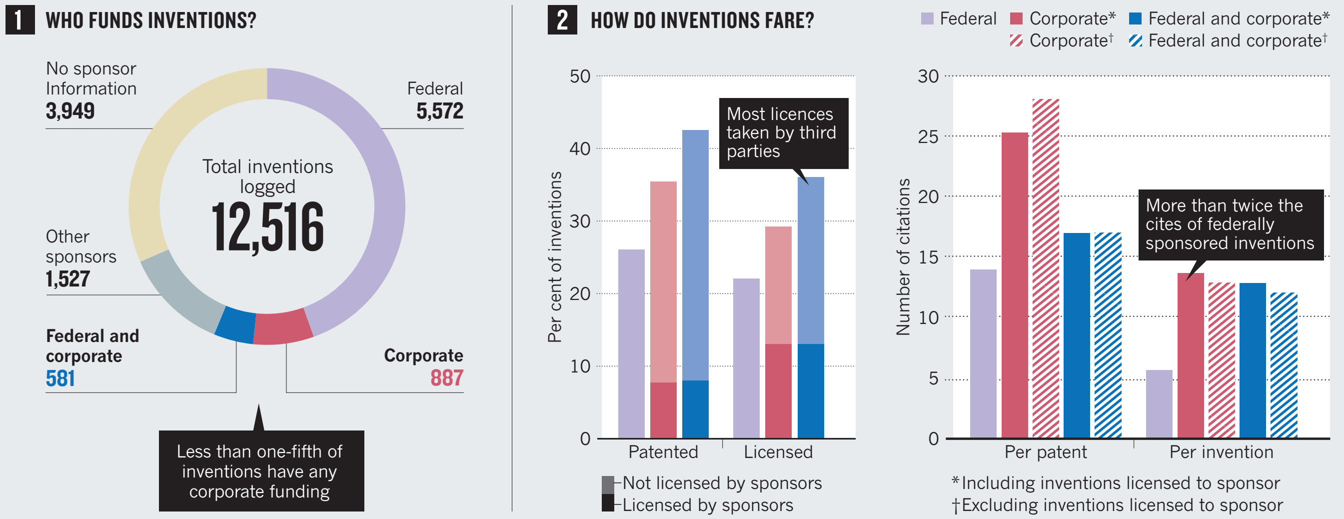
STOEBER, B. & LIEPMANN, D. *J. MICROELECTROMECH. SYST.* **14**, 472–479 (2005)/IEEE

**Microneedle fabrication, the subject of one of the most highly cited University of California patents.**

UC PATENT TRACKING SYSTEM DATABASE

## LICENSED AND CITED

Of the 12,516 inventions logged by technology-transfer offices of the University of California system between 1990 and 2005, inventions with only federal funding were less likely to be patented or licensed than those with corporate or corporate and federal funding, and had lower patent citation rates.




at the University of Chicago in Illinois suggests that corporations turn to universities to investigate areas outside their core strengths, investing in speculative science in the hope of finding profit opportunities[9].

In fact, Evans argues that corporations actually urge academics to explore further afield than they might otherwise. Although academics may act conservatively to gain acceptance of peers, papers and grant proposals, he writes[9], "industry partnerships draw high-status academics away from confirming established theories and towards speculation".

For example, the $500-million research grant from BP to the Berkeley-led consortium was intended to explore biofuels from cellulose in plants or crop residues, an area in which BP had virtually no expertise. In such cases, many resulting inventions might turn out to be informative to other researchers, but irrelevant to the firm's business strategy.

In such cases, other firms' subsequent work on an invention can be more valuable to sponsors than exclusive access. For example, preliminary work by Yongdong Liu, a PhD candidate at Berkeley, suggests that information-technology company IBM discloses innovations on the periphery of its expertise without patenting them, but often cites non-IBM patents building on the disclosed innovations. Similarly, some major drug companies contributed to the publicly funded Human Genome Project, reasoning that faster access to results would accelerate its ability to develop drugs, even if those results were openly available.

Acquiring intellectual property is not necessarily the prime focus of corporate sponsors. Companies also value sustained relationships with leading scientists and associated opportunities to identify and recruit talented employees. The University of California–Novartis agreement apparently generated no licences for the company, and Novartis representatives reportedly did not exert any apparent influence on the selection of projects it funded[10].

Joint federal–corporate sponsorship may stem from more-focused goals. We understand that they often arise from projects initiated by federal funding agencies, with corporate sponsors recruited to develop early, promising work into practical applications. For example, if a federally sponsored gene-screening programme finds an attractive drug target, corporations might support projects to screen drug candidates against that target. This kind of focus would explain why inventions in this category are the most likely to be licensed (even by third parties) but not more highly cited.

The large share of third-party licences suggests that the University of California successfully markets inventions and also negotiates agreements to keep corporations from locking them up unduly. This task is probably facilitated by the fact that many sponsoring firms seem to recognize that sharing exploratory research can be in their own interests.

To assess whether these findings generalize to other academic institutions, data from other research universities are needed. We advocate a project to pool similar data from a large sample of other research universities, with solid confidentiality safeguards, for empirical analysis. Such work could evaluate whether, for instance, groups of smaller or less research-oriented institutions would be better served by outsourcing to a single technology-transfer institution.

Universities setting up contracts with corporations need to be vigilant in their mission to generate and transfer knowledge, but they should not assume that companies are focused mainly on tying up intellectual property. Those that do will miss fruitful opportunities for collaboration with firms willing to fund projects from which many others will probably benefit. ■

**Brian D. Wright** *is professor of agricultural and resource economics at the University of California, Berkeley, USA.* **Kyriakos Drivas** *is a postdoctoral research economist at the Agricultural University of Athens, Greece, and a research fellow at the University of Piraeus, Greece.* **Zhen Lei** *is assistant professor of energy and environmental economics at the Pennsylvania State University in University Park, USA.* **Stephen A. Merrill** *directs the US National Academy of Sciences' Program on Science, Technology, and Economic Policy in Washington DC, USA.*
*e-mail: bwright@berkeley.edu*

1. President's Council of Advisors on Science and Technology *University–Private Sector Research Partnerships in the Innovation Ecosystem* (OSTP, 2008); available at http://go.nature.com/hilyum.
2. National Science Board *Science and Engineering Indicators 2014* 5–13 (NSF, 2014).
3. Washburn, J. *University, Inc: The Corporate Corruption of Higher Education* (Basic Books, 2005).
4. White, J. & Bero, L. A. *Stanford Law Policy Rev.* **21,** 105–133 (2010).
5. Kramer, D. *Physics Today* **61,** 20–22 (2008).
6. Litan, R., Mitchell, L. & Reedy, E. J. *Innov. Policy Econ.* **8,** 31–57 (2008).
7. Foley, H. C. *Res. Technol. Mgmt* **55**, 12–17 (2012).
8. Pressman, L. *et al. Nature Biotechnol.* **24,** 31–39 (2006).
9. Evans, J. *Am. J. Sociol.* **116,** 389–452 (2010).
10. Busch, L. *et al. External Review of the Collaborative Research Agreement between Novartis Agricultural Discovery Institute, Inc. and The Regents of the University of California* (Inst. Food and Agricultural Standards, Michigan State Univ., 2004); available at http://go.nature.com/sgoc1a.

# Innovative academic startups 2015

Brady Huggett

*Nature Biotechnology's* selection of academic spinouts ranked by amount of venture capital raised (**Table 1**) continues to be dominated by US ventures (7 of 10). Greater access to capital explains why the United States recorded 67 ventures of all types securing A rounds in 2015 (**Fig. 1**). Elsewhere in the world, the UK, China, Canada, Switzerland and France were the next most successful locations in terms of capitalization, with 10, 6, 5, 3 and 3 companies, respectively; the UK tops the list of average amount raised per round (**Table 2**).

**Table 1 Top ten series A rounds in 2015 for innovative academic startups (as of December 15)**

| Company | Amount raised (millions); date; investors | Scientific founders | Other founders | Technology |
|---|---|---|---|---|
| Gritstone Oncology (Emeryville, CA, USA) | $102; Oct. 20; Versant Ventures, The Column Group, Clarus Ventures, Frazier Healthcare, Redmile Group, Casdin Capital, Transformational Healthcare Opportunity | Tim Chan, Memorial Sloan Kettering Cancer Center; Mark Cobbold, Massachusetts General Hospital Cancer Center and Harvard Medical School; Graham Lord, King's College London; Naiyer Rizvi, Columbia University Medical Center; Jean-Charles Soria, South-Paris University | Andrew Allen, former CEO, Clovis Oncology | Identifying patient-derived, tumor-specific neoantigens for the development of individualized synthetic cancer vaccines |
| Neon Therapeutics (Cambridge, MA, USA) | $55; Oct. 1; Third Rock Ventures, Clal Biotechnology Industries, Access Industries | James Allison, The University of Texas MD Anderson Cancer Center; Nir Hacohen, Broad Institute; Eric Lander, Broad Institute; Robert Schreiber, Washington University; Ton Schumacher, The Netherlands Cancer Institute; Catherine Wu, Dana-Farber Cancer Institute | Ed Fritsch, formerly at Dana-Farber Cancer Institute and the Broad Institute of MIT and Harvard | Identifying patient-derived, tumor-specific neoantigens for the development of individualized and off-the-shelf cancer vaccines |
| Decibel Therapeutics (Cambridge, MA, USA) | $52; Oct. 15; Third Rock Ventures, SR One | M. Charles Liberman, Harvard Medical School; Gabriel Corfas, University of Michigan; Ulrich Müller, Scripps Research Institute; Albert Edge, Harvard Medical School | Not applicable | Novel therapies for hearing loss to modulate such targets as atonal homolog 1 (*Neuron* **77**, 58–69, 2013), repulsive guidance molecule A or its receptor, neogenin |
| Revolution Medicines (Redwood City, CA) | $45; Feb. 4; Third Rock Ventures | Martin Burke, University of Illinois at Urbana-Champaign | Mark Goldsmith, Third Rock; David Pompliano, Third Rock | Automated chemical synthesis of *N*-methyliminodiacetic acid-boronate containing intermediates (*Science* **347**, 1121–1226, 2015) to generate amphotericin analogs with antifungal activity (*N. Chem. Biol.* **11**, 481–487, 2015) |
| Semma Therapeutics (Cambridge, MA, USA) | $44; March 24; MPM Capital, Fidelity Biosciences, Arch Venture Partners, Medtronic | Doug Melton, Harvard; Felicia Pagliuca, previously a post-doctoral fellow in Melton's laboratory at Harvard | Robert Millman, MPM Capital; Jeff Imbaro, previously at Pursuit Solutions | Cell transplantation treatments for type I diabetes based on human pancreatic beta-like cells derived from ESC or iPSC (*Cell* **159**, 428–439, 2014) |
| Freeline Therapeutics (London) | $37.7; Dec. 10; Syncona Partners | Amit Nathwani, University College London | Christian Groendahl, Syncona | Pseudotyped, self-complementary adenovirus-associated virus subtype 8 vector expressing codon-optimized human factor IX for patients with hemophilia B (*N. Engl. J Med.* **365**, 2357–2365, 2011) |
| Metacrine (San Diego) | $36; Aug. 5; Arch Venture Partners, EcoR1 Capital, Polaris Partners, venBio | Ronald Evans, Salk Institute; Michael Downes, Salk Institute | Rich Heyman, previously CEO of Seragon | Protein sensitizers for insulin in type 2 diabetes and modulators of farnesoid-activated receptors for non-alcoholic steatohepatitis and other metabolic diseases |
| TherAchon (Biot, France) | $35; Sept. 30; OrbiMed Advisors, New Enterprise Associates, Inserm Transfert, Versant Ventures | Elvire Gouze, Inserm, University of Nice Sophia Antipolis | NA | Soluble human fibroblast growth factor (FGF) receptor 3 decoy prevents binding of FGF to mutant FGFR3 in achondroplasia (*Sci. Transl. Med.* **5**, 203ra124, 2013) |
| Kesios Therapeutics (London) | $28.8; Dec. 2; Imperial Innovations Group, SV Life Sciences, Abingworth | Guido Franzoso, Imperial College London; Menotti Ruvo, Istituto di Biostrutture e Bioimmagini of CNR; Laura Tornatore, Imperial College London | NA | Preclinical compounds targeting GADD45β/MKK7 complex downstream of NFkappaB (*Cancer Cell* **26**, 495–508, 2014) for multiple myeloma |
| Neurona Therapeutics (S. San Francisco, CA, USA) | $23.5; Dec. 1; The Column Group, Topspin Partners, private investors | Cory Nicholas, University of California, San Francisco (UCSF); Arnold Kriegstein, UCSF; Arturo Álvarez-Buylla, UCSF; John Rubenstein, UCSF | NA | Transplantation of human ESC- and iPSC-derived γ-aminobutyric acid–secreting (GABAergic) interneurons for epilepsy, neuropathic pain, spasticity and certain cognitive impairments and psychoses cells. |

Source: BCIQ: BioCentury Online Intelligence; company materials.

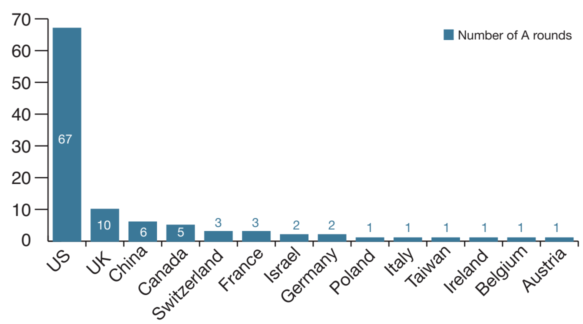



**Figure 1** Number of startups by country, 2015. A-2 rounds of undisclosed amounts left out. Source: BCIQ: BioCentury Online Intelligence.

**Table 2 Total and average series A rounds by country, 2015**

| Country (number of rounds) | Total amount raised ($ millions) | Average raised per round ($ millions) |
|---|---|---|
| UK (10) | 516.2 | 51.6[a] |
| Belgium (1) | 31.2 | 31.2 |
| US (67) | 1,482.2 | 22.1 |
| France (3) | 51.6 | 17.2 |
| China (6) | 89 | 14.8 |
| Canada (5) | 72.6 | 14.5 |
| Switzerland (3) | 43 | 14.3 |
| Germany (2) | 27.1 | 13.5 |
| Italy (1) | 11.2 | 11.2 |
| Taiwan (1) | 8 | 8 |

A-2 rounds, and rounds of undisclosed amounts left out. Source: BCIQ: BioCentury Online Intelligence. [a]Without outlier Immunocor, $21.8.

*Brady Huggett is business editor at Nature Biotechnology.*

Maksim Shmelov / Alamy

# Starting up and spinning out:

## The changing nature of partnerships between pharma and academia

By Wudan Yan



In 2005, two chemists decided to work together to develop libraries of cyclic peptides—strings of amino acids formed in the shape of a circle—that could be used to treat a range of infectious diseases, autoimmunity and cancer. Thanks to their flexibility, these cyclic peptides can access different parts of a protein target and so bind to targets often deemed 'undruggable'. Although these compounds sounded promising, computational chemist Matthew Jacobson at the University of California, San Francisco, and synthetic chemist Scott Lokey at the University of California, Santa Cruz, needed to understand how the compounds would work in cells before they could use them in animals or humans. Lokey and Jacobson's academic, theoretical work continued for the next six years.

During that interval, in 2009, pharmaceutical giant Pfizer signed a blanket agreement with QB3, an incubator for academic spinouts, with lab space spread across five sites in the Bay Area for early-stage biotech companies. When Spiros Liras, the former head of medicinal chemistry at Pfizer, read some of the papers Jacobson and Lokey produced, he thought that these peptides could be interesting to study and develop further. In 2011, Jacobson and Lokey signed an agreement with Pfizer to study these peptides in biological models to which neither of their labs previously had access. "Pfizer made their *in vitro* assays and animal pharmacokinetic tests available for our model systems," Lokey says. "Formerly, we had just been doing the best we could with simple cell-free assays for permeability because we didn't have the other resources at all." The scientists subsequently published a paper detailing the results (*Nat. Chem. Biol.* **7**, 810–817, 2011).

After the paper had been published, however, Jacobson and Lokey did not immediately think to start a company. "I didn't know the first thing about starting a company," Lokey says. "It seemed like such a massive undertaking, but QB3 has provided a lot of support and helped us walk through the initial stages." Through QB3, the chemists were introduced to David Earp, a research scientist with business experience, and ultimately, a company called Circle Pharmaceuticals was established as a private entity based in San Francisco in May 2013. Earp is now the CEO and president of Circle. Although Jacobson's and Lokey's partnership with Pfizer ended officially in April 2015, the company still funds Circle's work. The continued collaboration with Pfizer and additional seed funding from San Francisco–based venture-capital firm Mission Bay Capital have allowed Circle to further the preclinical development of its therapeutics (*Medchemcomm* **3**, 1282–1289, 2012; *Curr. Top. Med. Chem.* **13**, 821–836, 2013). Circle has now set up shop in the South of Market district in San Francisco, a major start-up hub.

The creation of startups such as Circle from collaborations between academia and industry is the "new exciting trend" in the field, according to Doug Crawford, associate director of QB3. "These efforts began in earnest around 2011, although there were earlier efforts, as well," he says. Not only are startups spinning out of existing collaborations, but pharmaceutical companies have also been investing and partnering more with new, early-stage startups, particularly in the last five or so years, with the ultimate goal of commercializing tools and compounds to advance the field of medicine.

## A history of collaboration

The enactment of the US Bayh-Dole Act in 1980 enabled publicly funded academic medical institutions to interact with commercial entities to collaborate, opening up more partnerships between industry and academia. Before Bayh-Dole, university labs had served primarily as centers to investigate basic research questions, with little concern for commercial application. Bayh-Dole enabled research findings to be translated more rapidly into clinical use; over the past 30 years, more than 150 FDA-approved vaccines, drugs and new indications for existing drugs have been discovered through studies carried out in research institutions (*N. Engl. J. Med.* **364**, 535–541, 2011).

Collaborations between industry and academia have taken a variety of forms that have evolved over the years. The first era of these partnerships, from the early 1990s to 2007, primarily utilized the 'fee for service' model, in which academic institutions were paid a fee to perform experiments that would further the goals of their industry partner. For example, the University of California, Berkeley, signed a 5-year, $25-million contract in 1998 with Novartis Agricultural Discovery Institute, a research arm of Novartis, to help the company with its research projects. Eight years later, in 2006, the Scripps Research Institute in La Jolla, California, signed a similar five-year agreement with Pfizer that gave Scripps a $100-million cash infusion and Pfizer the rights to license approximately half of the discoveries from the partnership. However, the pharma partners were not interested in executing any licenses from either of those collaborations, and, according to Sarah Cairns-Smith, a senior partner at the Boston Consulting Group, "both of these partnerships were big failures."

In the years since, however, both pharmaceutical companies and academic institutions began to rethink the nature of pharma-academia partnerships. The research interdependence between the two parties allowed pharma to diversify their portfolios into unmet medical needs without having to find the capabilities to do so in-house, and academia reaped benefits from the expertise in drug development and resources provided by industry. With pharma facing patent cliffs in the past decade and an almost 20% decline in US National Institutes of Health research grants from 2003 to 2014—adjusted for inflation—both parties recognized the need for collaboration. These incentives have inspired pharmaceutical companies to gradually move away from the old "fee for service" model, such as the Novartis-Berkeley or Pfizer-Scripps deals.

"In these fee-for-service models of partnership, pharma would own all the results of the work done by academics," says Juan Carlos López, head of the Academic Research and Collaborations group at Roche in New York (López was formerly chief editor of *Nature Medicine*). "But academics no longer want to be a part of a partnership in which they do the experiment, get paid by the company, receive money and then give all the findings back to the company. Academic institutions now want to benefit from the rewards from the research they've undertaken."

"We're making sure we're protecting the inventions as early as possible to get patent rights."

The 'second era' of partnerships now considers how best to align the interests of academics with those of pharmaceutical companies. "We now have a new way of doing business with academics," López says. "Before it was: let's buy or license from academia and then develop it ourselves. Now, what's new is that we're interacting with academic institutions to co-develop something. We also want both parties to share the rewards."

Alan Rigby, vice president of Eli Lilly's antibody drug conjugate program in New York, agrees. "Academic researchers are more interested in retaining rights to their work. They're less reluctant to sell, so both academics and industry partners are evaluating options to keep both parties involved." Industry partners are now looking to 'alliance managers' within the academy, such as those people who work in technology transfer offices of universities and academic medical centers, to find research teams that can complement their own existing programs. "These partnerships aren't all about solving the innovation shortfall in pharma anymore," Crawford says. "Rather, productivity is a function of reasonable interest on both sides in the science."

**The growth of patents and startups at US hospital and university tech transfer offices**

| Year | Total licenses and options executed | Issued US patents | Start-ups formed |
|---|---|---|---|
| 1999 | 3650 | 3477 | 292 |
| 2000 | 4004 | 3548 | 386 |
| 2001 | 3657 | 3545 | 424 |
| 2002 | 4247 | 3489 | 401 |
| 2003 | 4473 | 3926 | 374 |
| 2004 | 4758 | 3667 | 462 |
| 2005 | 4897 | 3272 | 451 |
| 2006 | 4947 | 3245 | 554 |
| 2007 | 5094 | 3618 | 555 |
| 2008 | 5123 | 3289 | 595 |
| 2009 | 5321 | 3415 | 596 |
| 2010 | 5356 | 4465 | 651 |
| 2011 | 6037 | 4699 | 671 |
| 2012 | 6360 | 5150 | 705 |
| 2013 | 6549 | 5709 | 818 |

SOURCE: AUTM

## Internal changes

To ensure that these partnerships are productive, both academic institutions and industry partners have created internal teams on their individual sides to help to align the interests and skill sets in the collaborations that form. Technology transfer offices in academic institutions are now more eager to help academics to set up their own companies by performing outreach into the private sector.

Between 1990 and 2007, startups began to emerge from universities, which were realizing that, with certain inventions that were based on early-stage or platform technology, it would be difficult to get an established company interested in pursuing the ideas. These technologies could be developed further, and with more data, investors or larger companies might consider funding, partnering with

or acquiring the startup. During this time frame, state governments and economic development experts began to see university startups as a way to create local jobs.

Since 2007, technology transfer offices and universities started to learn how to support startups more effectively and efficiently. “Creating and licensing technology to a startup is much more complicated than doing a ‘traditional’ license to an established company,” says Fred Reinhart, president of the Association of University Technology Managers (AUTM) near Chicago. “Most schools use a different set of license terms and different negotiation approaches, which are customized to the needs of a nascent enterprise. In addition, universities are becoming more aware that to be more successful at creating startups, you need to create an environment that encourages and supports entrepreneurship on campus, such as clear policies and sensible procedures for handling inventions destined for licensing to a startup, access to qualified management and access to financial resources.”

There are indications that these efforts are already bearing fruit. AUTM has been tracking statistics on academic tech transfer since 1991. Between 1999 and 2003, licenses and options executed across all industries of academic technology in the United States increased by 22%. The total number of licenses and options executed increased by another 7% from 2003 to 2008, and by 23% from 2009 to 2013. The number of patents issued rose 12% from 1999 to 2003, decreased 10% from 2004 to 2008 and then rebounded by nearly 70% from 2009 to 2013. 596 startups had spun out of universities and academic medical centers in 2009; by 2013, that number had increased to 818. Today, there are more than 4,200 operational startups that have spun out of academic work in the US.


Patti McConville / Alamy

**Rising trend:** The Alexandria Center for Life Science in New York.



For example, although QB3 was started in 2000 as part of a state initiative to grow the California economy, by 2004 it had started serving as an alliance manager within the University of California system to help pharmaceutical and biotechnology companies to partner with academic startups. QB3 has helped not only to find external partners for early-stage spinouts but also to incubate these teams. “Pharmaceutical companies are turning to QB3 not to connect with the academic world, but with the startup world. It’s more efficient with returning ideas of products with interest to us,” Crawford says. QB3 has housed 138 programs since 2006, currently rents space to 80 companies and has graduated more than 40 companies since its inception.

Universities outside startup hubs such as the Bay Area are also revamping their technology transfer efforts to best support their researchers’ products and inventions. “Researchers are becoming more entrepreneurial,” says Sadhana Chitale, a technology transfer officer at New York University (NYU). Most startups emerge from postdoctoral fellows or graduate students who want to take projects from the lab one step further. More than 70 startups from NYU have formed since 2012. “In the last 15 years, tech transfer offices have been more proactive in developing in-house processes, like providing gap funds and other internal resources before potential projects can be taken to the next level by receiving venture-capital funding, or partnering with pharmaceutical companies.”

Technology transfer offices are also taking steps to ensure that academic scientists have proper ownership over their work. “We’re making sure we’re protecting the inventions as early as possible to get patent rights,” says Reinhart. “We’re also working with investors to make license agreements that will allow startups to succeed.”

“Pharmaceutical companies are more interested in driving the creation of a deliverable, which is more in line with what startups are doing.”

Around 2010, a development occurred that helped more startups to emerge: pharmaceutical companies started to build their own venture-capital arms to provide seed money for startups that may arise from former partnerships or for companies that do work that is complementary to their priorities. “With the economic downturn, pharma saw a decrease in biotech spending from traditional life-science venture-capital firms, so they initiated their own venture funds,” Earp says. “Academics are appropriately focused on understanding basic scientific processes, whereas pharmaceutical companies are more interested in driving the creation of a deliverable, which is more in line with what startups are doing.”

Total venture-capital money spent on biotechnology was around \$4 billion aggregate dollars in 2012 and is projected to hit \$7 billion in 2015, according to Sam Kulkarni, a partner at McKinsey & Company who is based in Silicon Valley.

### Spinning out and starting up

The creation of venture-capital arms in pharmaceutical companies during the past five years, coupled with the changing efforts and priorities of university technology transfer offices by providing internal funds, have enabled the creation of more productive partnerships and startups, according to Reinhart. The creation of early-stage life-science startups, such as Circle Pharmaceuticals, serves as an example of how the nature of these partnerships between industry and academics is changing with the interests of both parties in mind: industry partners do not have to conduct experiments in house, where R&D funding is also low, and academics would have ownership over their research.

The inception of 4D Molecular Therapeutics is another example of how startups are spinning out of academia and partnering with pharma. David Schaffer,

a chemical engineer at the University of California, Berkeley, had been working on developing gene therapy techniques since he joined the university in 1999. "I had been thinking about starting a company since around 2007, but the economic crash at the time made it clear that the timing wasn't good for a startup. In the first eight to nine years of this work, I don't think biotech was ready for gene therapy," Schaffer says.

After Schaffer met David Kirn, a physician-scientist and entrepreneur, through QB3 in 2012, the two started to talk more seriously about forming a company. They founded 4D Molecular Therapeutics (4DMT) in 2012, with the help of QB3. 4DMT officially incorporated in September 2013. It is currently the largest company to be housed in QB3, with nine full-time scientists and five legal consultants. With the help of QB3, 4DMT signed a deal with Roche on 27 April to generate and optimize adeno-associated viruses to treat several ocular diseases. 4DMT also has ongoing collaborations with Amsterdam-based UniQure and with Applied Genetic Technologies Corporation (AGTC) in Gainesville, Florida.

However, not all successful partnerships result in the creation of a startup. France-based pharmaceutical company Ipsen announced a collaboration with Harvard University in 2013; the agreement was renewed in 2015. Ipsen has been partnering with microbiologist Min Dong at Harvard Medical School, who has been developing botulinum toxins to treat neuromuscular conditions since joining Harvard in 2009. Because the goal of the current collaboration is to develop a product for Ipsen based on a filed patent that Ipsen licensed, the initiation of a startup from this agreement does not make sense; the creation of a company would be outside the scope of the collaboration.

According to Reinhart, companies that spin out from academic work set up shop in their own states approximately 75% of the time. "If these companies succeed, then economic activity increases in the same state," Reinhart says. Life-science spinouts are most prevalent in cities where there is a high concentration of academic medical centers with a mature biotech and venture-capital community, such as San Francisco and San Diego in California and Cambridge in Massachusetts.

Cities such as New York, however, are also making strides by establishing lab spaces. In 2010, Accelerator Corp., a biotech investment and management firm, worked with Alexandria Real Estate Equities, which owns and develops life-science facilities across the US, to open the Alexandria Center for Life Science, a glassy building on the east side of Manhattan. The center would provide laboratory space to incubate local spinouts, in addition to housing branches of pharma companies such as Eli Lilly, Roche and Pfizer's Center for Therapeutic Innovation. Harlem Biospaces, a biotech incubator tucked away in the northwest corner of Manhattan and founded with financial support from the New York City Economic Development Corporation, opened in November 2013. It has the space and resources to incubate up to 24 early-stage life-science companies. Despite the resources that are available for startups to grow out from academic labs, the current cost of renting lab space in New York is still higher than what most early-stage startups can afford. "The majority of companies that spin out from NYU set up shop elsewhere," Chitale says.

To mitigate the high costs and taxes of operating in New York State, New York governor Andrew Cuomo created an initiative called START-UP NY for companies located on or near eligible university or college campuses to operate tax-free for ten years. Furthermore, the New York City Council introduced legislation to provide a refundable tax credit up to $250,000 for small (fewer than 100 full-time employees) biotech companies based in New York.

"Academic researchers are more interested in retaining rights to their work. They're less reluctant to sell, so both academics and industry partners are evaluating options to keep both parties involved."

### The waiting game

Some partnerships, particularly those with the goal of bringing a drug into the clinic, have yet to bear fruit given the amount of time it takes for a compound to make it through development. "If we measure the success of those partnerships by evaluating how many drugs are brought to market that have arrived from us doing work with external partners, then it's too early to say," López says. "What we're able to say in the shorter term is how these drugs are advancing in the pipeline." Currently, a third of Roche's pipeline has originated from work with external partners—this value has remained about the same over the past 15 years.

However, for now, consultants think that these partnerships are here to stay. "Is the bubble about to burst?" asks Jon Duane, director of pharmaceutical and medical products at McKinsey & Company, who is based in Silicon Valley. "I imagine these partnerships will continue unless there's a major disruption. In the last two to three years, we've seen more venture capital flow back into biotech."

For startups that have already spun out, such as Circle, the scientists think that the former and existing collaborations with pharma companies have been important for their work. "It's easier for a [pharmaceutical company] to negotiate rights with a small company as compared to negotiating rights with a lab in a research institute," Earp says. "Working with Pfizer has been absolutely foundational to creating Circle and getting it off the ground. In the meantime, we at Circle have also been working on targets of interest to Circle, so the initial partnership is growing our own company, as well."

The work that Circle does in developing cyclic peptides as drugs is an area that can benefit from the complementary expertise and skill sets of industry and academic researchers, according to Joshua Kritzer, a chemical biologist at Tufts University in Medford, Massachusetts.

"Medicinal chemists have long appreciated that cyclic peptides can be potent and useful drugs. There's an abundance of therapeutic natural compounds in this class, and the rest are hiding in plain sight," Kritzer says. "It would take very long for pharmaceutical companies to discover and develop these compounds in house. That's where Lokey's and Jacobson's work comes in—they've developed ways to make cyclic peptides more drug-like. Although it's possible for academics to develop a lead compound of this type into a therapy, the drug discovery process would be much longer if academic scientists just worked amongst themselves."

*Wudan Yan is a freelance science journalist and former news intern for* **Nature Medicine.**

# Aligning needs

Dennis Ford & William Kohlbrenner

**The best way for aspiring entrepreneurs to achieve their financing goals is to understand what investors and partners want.**

If you are an aspiring entrepreneur spinning out a new biomedical technology or launching a biotech startup, one of your first tasks is to understand the funding process and how to tackle it. Five years ago, the accepted investment path was to write a proposal for Small Business Innovation Research (SBIR) grants, hit up a list of friends and family and canvass the local regional angel groups for 'seed' funding. After this first wave of funding, the next money was expected to come from venture capital (VC) entities or through collaborations with partnering companies.

What you need to know is that the world has changed, and the investor landscape has morphed. New types of financiers have entered this space, many hoping to accelerate the translation of basic research (**Fig. 1**). Certain US states now have programs that provide seed funding, with the goal of increasing local startup activity. Some support life science activity in general, such as Massachusetts Life Science Center or New Jersey Economic Development Authority's Technology & Life Sciences programs, whereas others focus on particular local strengths such as the California Institute of Regenerative Medicine or Cancer Prevention & Research Institute of Texas. Many US universities are establishing seed funds to benefit their own academic entrepreneurs, following on from similar pioneering efforts in Europe. Research institutions are now also launching commercialization funds and incubators, such as Texas Medical Center's TMCx. Corporate VC efforts have grown, with many pharma, medtech and information technology companies willing to back biotech startups. Family offices of high-net-worth individuals may also make investments in promising companies—whether it be for personal reasons or market opportunity. And there are a growing number of venture philanthropists, patient groups and foundations open to supporting basic research and seed-stage ventures in specific areas.

*Dennis Ford (Founder & CEO) & William Kohlbrenner (CSO) are at Life Science Nation, Boston, Massachusetts, USA.*
*e-mail: dford@lifesciencenation.com*

As CEO or founder of a startup, your goal should be to understand the motivations and desires of all these investor types (**Table 1**), how they operate and how to approach them. You should learn the individual mandates of each investor, as well as a defined set of 'knockout' factors that could eliminate your company from further consideration (**Box 1**). This article will help you understand what buyers are looking for before you approach them.

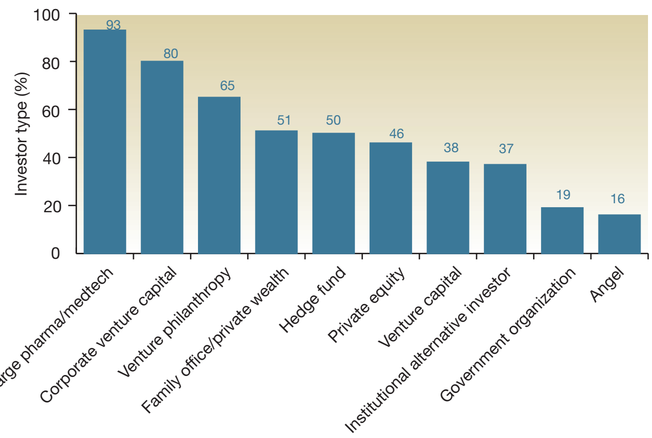



**Figure 1** Percentages of each investor type that are seeking opportunities globally. Source: LSN Investor Platform, as of 1 October 2015.

## The buyer's mind

Investors are highly specialized and usually short on time. Remember, your success in attracting funding may correlate more with your technology's development stage, level of risk and extent of validation than with anything else. On the other hand, early-stage opportunities often have a lower cost of buy-in compared with a de-risked but later-stage asset. This allows investors to make multiple bets on several early-stage opportunities instead of just a few, more-expensive ones.

In what follows, we present several key criteria that investors use to quickly evaluate early-stage opportunities. Take heed. Any presentation you make to a group of investors should address each of these criteria.

**Management team.** Investors want an experienced management team in place before investing. The reasoning, of course, is that seasoned entrepreneurs are thought to have a higher probability of again finding success. Thus, include the names and credentials of the management team. If your core group is young and inexperienced, seek out SBIR grants and collaborate with knowledgeable technical advisors and seasoned business colleagues, as both will help you establish credibility and raise funds. Take the time to recruit older experienced mentors that can fill the business-side holes until the team is intact. This shows a rudimentary understanding to the potential investors, which they need to see.

**Unmet need.** Investors love products that can satisfy a significant unmet need, as these have the possibility to transform current standards of care and can command a high price tag. Similarly, programs aimed at alleviating so-called orphan diseases receive a great deal of attention from pharmaceutical companies, investors and venture philanthropists because

## Box 1 A knockout game

Investors are inundated with entrepreneurs soliciting their help, advice and capital. Their goal is to swiftly get compelling opportunities on the table and remove ones that are not a fit.

Below is a list of other reasons why investors can easily disregard your pitch.

- Out of investment scope
- Too early for major investment (lacks validation, too much risk)
- Lacks sufficient IP coverage
- Lack of confidence in management team
- Expensive new product that lacks compelling rationale for replacing lower cost, current standard of care
- New product that brings modest incremental improvements over currently approved products that adequately address medical need.

of their lower regulatory hurdles, their exclusive market or data, and the commercial potential to expand to broader indications after approval. As more and more groups have emerged syndicating support around specific diseases, increased funding is available to find solutions for rare and neglected diseases.

**Market fit.** Early-stage opportunities targeting established markets (the United States, the European Union and Japan) are likely to attract more investor interest than opportunities focused on emerging markets. However, there is a subset of investors focused on the emerging market space with the goal of supporting initiatives that address global problems in infectious diseases and other areas.

Early-stage investors are aware that there is enormous pressure to contain and/or reduce the cost of healthcare globally, with institutional (payer) gatekeepers or governments aiming to control access to new therapies and technologies. It will be critical to develop a compelling rationale that justifies a switch from the current standard of care to your premium-priced product. Investors will assume that market uptake of products that bring only incremental improvements may be limited, and they may, therefore, be unwilling to invest in your company.

When speaking to investors about the market for your potential product, don't project

**Table 1 Overview of investor classes**

| Investor class | Profile | Investment goal |
|---|---|---|
| VC | VC funds are very selective and establish large funds that are used for investing in a portfolio of companies that they view as having a high probability of success accompanied by a rapid increase in valuation. They often prefer working with experienced entrepreneurs. There are VC funds in the early-stage space; many now focus more on established companies than startups. | VC funds want to invest in promising early-stage companies that have strong potential for an initial public offering (IPO) or that can be sold to a strategic partner, allowing an early investor exit with high return on investment (ROI). |
| Private equity (PE) | PE funds typically invest in market-stage companies generating revenues, rather than startups. However, some PE funds (such as TPG Biotech, Yuanta Asia Investment and GTCR Golder Rauner) are open to exploring select early-stage opportunities. A substantial investment is made to buy the company, which is then restructured and sold at a profit. | Short-term ROI based on rapid sale of restructured asset. |
| Angel investor | High-net-worth individuals, with an interest in a particular type of product, service or industry. These have traditionally been the dominant go-to group for seed funding of startups. Many are successful entrepreneurs themselves. May join networks to increase size of investment pool. | Investor focus is on companies in the earliest startup stage with the goal of funding promising technologies they view as having high potential value. |
| Venture philanthropy | Foundations, nonprofits and patient advocacy groups are typically focused on specific disease areas that provide grants for basic academic research and support the development of drugs through venture investments. | Accelerate the development of cures for specific diseases. Some philanthropic groups use an evergreen structure in which ROI is returned to the fund for future work. In other cases it is nondilutive financing. |
| Hedge fund | As yet, only a few active in the early-stage life sciences. Pool of capital from a number of investors, and that is invested in securities and other instruments. Some hedge funds are open to exploring select early-stage opportunities. In such cases, more likely to pool funds with other entities. | Investment strategies aim to achieve a positive ROI regardless of whether markets are rising or falling. |
| Big pharma/ biotech/medtech | Pharma, biotech and medtech giants devote substantial resources to identifying development-stage or marketed products that can be introduced into their product portfolios through exclusive in-licensing or company acquisition. | Obtain exclusive access to products that can be introduced to the market over the near term. |
| Corporate VC | Many large companies allocate funds for investing in early-stage technologies or products that align with their strategic goals. In the life sciences, corporate VC funds typically act as co-investors in financings. | Corporations seek early access to opportunities that can enhance their pipelines over the long term. Achieving a high ROI is not necessarily a major investment goal. They do want home runs, but primarily focus on building a strategically significant portfolio. |
| Family office/ private wealth | Represent the collective estate and assets of ultra-high-net-worth individuals. Generally maintain a low profile but have large amounts of capital, a sophisticated institutional investing approach and a long-term outlook. May also have an interest in philanthropy. | Investing with the goal of achieving significant ROI, but with a long-term outlook. Some family offices may want to invest early to help stack the odds of helping find a cure for a family disease or malady. |
| Institutional alternative investor | Includes financial institutions, pension and endowment funds and other entities that are seeking to diversify their holdings and are open to expanding their portfolios to include high-risk, high-return opportunities. | High ROI from key investments over the long term that enhance portfolio value. |
| Government agencies and universities | Government agencies in the USA provide grants to startup companies though the SBIR/Small Business Technology Transfer (STTR) program. Some states have established programs that fund startups in the life science area. More recently universities have been providing seed funding to help entrepreneurs bring their technologies out of the laboratory. | The aim of these programs is to help entrepreneurs commercialize promising basic academic research. Typically involves nondilutive funding (government doesn't own part of the company). |

## Box 2 Expanding your virtual rolodex

Well-funded investors are not looking for just a single deal; they are seeking to build a portfolio of investments, and that requires substantial 'deal flow', meaning opportunities are continually evaluated, vetted, prioritized—with the best ultimately funded. To achieve this, many investors attend select conferences, extensive networking activities and industry events—all of which increases their odds of finding the best deals and most compatible investment partners.

Many investors have a web presence, and if a firm is looking for in-bound deal-flow, their website should suggest an initial point of contact, or provide staff profiles that allow an entrepreneur to identify the most relevant person to approach. Take the time to research the investor's portfolio and, if you think you're a close fit for the firm's interests, you can attempt to start a dialog. LinkedIn is also a useful tool for finding staff at investment firms who have experience in your particular area of science.

However, certain investors may not be interested in in-bound deal-flow at all. For example, some family offices prefer to operate in 'stealth mode', and they source opportunities through proprietary networks or preferred syndication partners. More recently, some investors (particularly major life science VC funds) are pursuing 'build-to-buy' investment approaches, in which the investor sources IP directly from a university or research institute and builds an executive team to take the asset towards commercialization, usually spinning out an LLC entity to hold the IP. Some of these investors do not invest in external entrepreneurs at all.



commercial success simply because you are in a particular multibillion-dollar therapeutic space, such as oncology. Rather, detail how you can successfully fit into that market.

**Development stage.** Because of the poor returns achieved by many early-stage VC funds over the past decade, many VC funds have shifted their focus to later-stage opportunities, leaving only a small cadre of boutique VC funds catering to the early-stage enterprise. There is no sense in approaching uninterested investors, so you will need to research your investors and make sure that they invest in companies at your development stage and sector.

**Validation.** Achieving validation of your product raises the value of your company's asset by decreasing risk. Although there may be multiple technical validation steps involved in product development, true validation is when your product performs as designed in a clinical setting; even getting a positive signal in a small phase 2 trial can be enough to boost an asset's value. Investors will want to see in what ways your product has been validated.

**Product differentiation.** New healthcare products face a competitive, highly regulated market with multiple barriers to entry. You will need to demonstrate the potential to clearly differentiate your product in the target market. That means having a detailed understanding of the commercial landscape you hope to enter—including already marketed products and those in the pipelines of competitors—and then articulating why your product will enjoy meaningful uptake. Doing this will increase your odds of scoring funds.

**Intellectual property.** Although filing invention disclosures and patent applications can be distracting for bench researchers, it is a front-and-center priority for the scientist entrepreneur. Investors will expect that you have protected your technology with a proprietary intellectual property (IP) position, so be prepared to demonstrate your right to operate in your space. This could mean acquiring an exclusive option to existing IP (university technology or otherwise) or filing relevant patent applications at the earliest stages of company formation. Anticipate an ongoing investment of time and money to strengthen your IP portfolio as the company moves forward. This involves a multiyear process working with the US Patent and Trademark Office and other agencies to get key claims issued and filing additional applications related to novel uses and manufacturing. In addition, you may need foreign filings for protection in global markets. This will require access to experienced (and expensive) legal services.

**Strategic alliances.** Getting in the door of an investment house is difficult enough; getting access to decision makers in established biotech or pharmaceutical corporations may seem even more daunting. But many of these companies are increasingly looking to partner with academics early in the discovery and development process (the move from R&D to search and development). And if you can find a partner in industry willing to back your work not only through research funding and operating capital, but also validation of your technology or molecule, investors will take note. They want to see if you have partnerships in place, as both the funding from these deals and the validation they bring can decrease risk for an investor. Decreasing risk can also come from US National Institutes of Health funding. It can come from a partnership with a foundation, patient group or philanthropy helping to move your technology through the development cycle. All of these external collaborations are indicators that others have gauged your technology and found it worth an investment of time and resources.

**Innovation plus.** You need more than cool science. If you are in the early stages of forming your company and talking to angel investors, then the technical innovations you bring should be the emphasis of those discussions. However, if your company has progressed further, and you are contemplating a major funding round, it's not enough to say you have an innovative platform. Your innovative science should be clearly reflected in an asset that sets you apart from the current market.

**Marketing materials.** This might seem like a small thing—your handouts, your PowerPoint

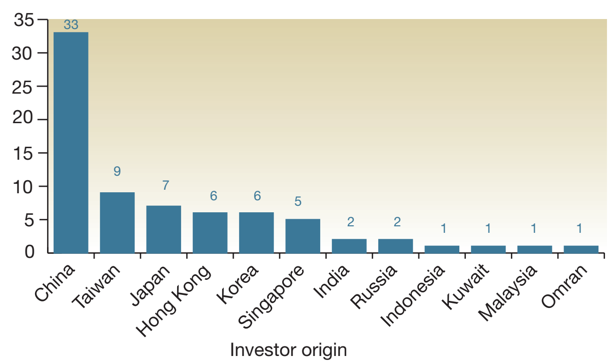



**Figure 2** Beyond home shores. Number of Asian investors looking for global opportunities by country. Source: LSN Investor Platform | Data as of October 1st 2015

presentations and your web presence. But those are often where investors get their first impressions of both you and your company. Your financing success will in large measure be determined by how skillfully you put together these fund-raising tools, and how you present yourself. Your company may be screened based in part on the quality of your marketing materials.

### You and your company

Life science entrepreneurs who started at the research bench often have little training or experience in marketing and indeed in how to market themselves, let alone a company. The best way we can think of to improve your odds of fund-raising success is: apply marketing and sales concepts to a fund-raising campaign.

**The list.** Your first step is to generate a list of investors that fit your company's stage and sector (**Box 2**). We have covered this previously (*Nat. Biotechnol.* **32**, 15–23, 2014), but keep in mind there are ~10,000 investors around the globe, and 95% of them probably are not a good fit for you. Doing your homework should drop that number to 300–500 investors. The goal from there is to do a first pass and get the list chopped down to a 100 or so, and then do more vetting and qualifying until it's down to 30–50. Once you get to that 30–50, use meetings and phone calls to find out who currently has the ability to give and interest in allocating funds, and reduce it further to 8–12 targets. The next step is reducing it to 3–4 investors who are serious, seeking opportunities and willing to pull the trigger.

This entire process—it's called campaign management—should take you 9–18 months (it's true that some CEOs can raise capital in 6–9 months, but they are the exception rather than the rule). This will take a lot of grunt research, but there are many low-cost cloud-based tools that can help whittle your list down. Automating the tasks of campaign management is key and allows you to track the tasks and interactions associated with each targeted investor.

**Branding.** Consider that there are many, many opportunities out there for early-stage life science investors. Financiers routinely state that they get hundreds of solicitations coming over the transom per week. As a result, they have gotten quite efficient in how they parse solicitations. They will judge you on (among other things) professionalism, presentation, intelligence and attitude.

Investors will expect you to have done your homework and understand their firm. Investors want to see cogent and lucid presentations that have more than a modicum of forethought and understanding of the task at hand. Branding and messaging that appears nonlinear, helter-skelter, too simple or too complex won't impress anyone. An experienced investor might do a cursory parsing of a solicitation in a couple of seconds and a first scan in a couple of minutes, so the easier you make it for them to understand the opportunity that you provide, the better chance you have of receiving a return phone call or e-mail.

**Referrals.** We cannot emphasize this enough. Referrals can be wondrous door openers. But remember, you will be part of an unfavorable situation if you are referred to an investor and end up not being a fit for their mandate. Investors don't make capital investments simply because of connections, so do not set up meetings simply because you can. Remember that the global life science universe is a relatively small one, and players, from the discovery phase to preclinical/clinical laboratories, right though to commercialization, can overlap as careers morph and companies progress. The people you meet now might in ten years be in new positions. There are good referrals and bad referrals—take the time to know the difference.

**Methodology.** How is your financing campaign organized? Whether you're deploying an in-house business development team, or working with an investment bank or third-party marketer, the staff that executes your campaign needs to have an efficient and reliable means of organizing and storing all the relevant data points. We believe that using a cloud-based customer relationship management (CRM) system to organize your campaign is essential. There are many CRM systems available at a low cost (typically $5–20 per user per month); we use http://www.Salesforce.com (we have no affiliation with Salesforce).

These programs allow you to import your investor target list information from third-party sources (usually through Microsoft Excel), which will create an account page and/or profile for each investor, which then serves as a home for tracking. It offers customized fields for data points related to each investor, such as date of last e-mail, follow up, last voice mail and more.

This allows the campaign team to organize their efforts; indeed, Salesforce.com allows users to automatically track their e-mails to potential investors using the "Email to Salesforce" setting. One benefit of the CRM system is that it can be used as the 'source of truth'; rather than team members spreading information across e-mails, Excel spreadsheets and meeting notes, all information is centralized in the CRM system.,

The information in these systems is useful for managing an in-house fund-raising team, but it's also essential if you are using

### Box 3 Looking to the East

Life science investment used to be a local affair. And indeed, many VCs prefer companies to be based locally so that attending board meetings or catching up with management does not involve flights around the globe. However, in recent years, an increasing number of early-stage investors—particularly investors outside of the traditional hubs like Boston and the Bay Area—are now thinking globally in their deal sourcing and investment efforts. Innovation knows no geographical restriction, and the investment community is acutely aware of this fact. Of the ~950 investors interviewed by our company Life Science Nation, 45% are open to investing globally or across multiple continents.

Having interviewed more than 100 Asian-based investors, 75% of them are open to making investments or to license technologies from outside of Asia. There are several factors leading to this trend. The first is that the current market for life science companies, particularly in the United States, is an attractive one. With the possibility of an IPO for the strongest companies still available, there is potential for a profitable exit. In addition, compared with Asian companies, US and EU life science companies have a stronger support ecosystem for innovation—from strong academic institutions performing discovery research, service providers able to assist in the drug design, development and clinical testing, as well as a larger pool of experienced entrepreneurs well versed in and willing to take new technologies to the market.

Finally, due to the large amount of interested capital in Asia and relatively few investable life science companies, the laws of supply and demand take hold and can drive up the price of Asian deals, making them less attractive than looking for overseas assets. The Asian-based investors we've spoken with generally are most interested in investing with the option to purchase distribution rights in their local geographies, rather than obtaining exclusive global rights to the asset. These groups tend to have strong connections with manufacturers and distribution channels in their regions and can serve as excellent partners in capturing market share in Asia for your product.

a third-party marketer. We've spoken to life science executives who paid a retainer to an investment bank, but had little insight on what this third party was doing. It's important to know which investors are being contacted, and how frequently. You'll want to see the message a broker sends to investors. CEOs typically hear from their fund-raising partner only when an investor meeting has been booked, but there is so much more to fund-raising than that. A program like Salesforce opens a window on the process.

Lack of adequate follow-up is the number one reason campaigns are not successful. Meeting an interested investor is similar to starting a conversation, and a conversation turns into a relationship only if it is monitored, nurtured and continued. Both parties can get busy, so you will need to make sure someone from your end steps into the breach and feeds this nascent interaction.

## Conclusions

It's important to manage your expectations regarding fund-raising. There is a hierarchy involved, with high-profile academic entrepreneurs at the top, who have multiple successes in building startups and who have relationships with top investors. For these people, a few phone calls may be all that is required to launch a company. However, for most neophyte entrepreneurs, the process will take substantially more time and effort. You must network at scientific meetings and partnering conferences, and keep in mind that many investors are increasingly looking to invest globally, especially investors based in Asia (**Box 3** and **Fig. 2**).

This is not for everyone. So before you start down the path of launching a new enterprise, be sure you appreciate and understand the challenges that any new entrepreneur faces. These challenges can be mastered. The trick is vetting your technology with a network of experts, coalescing a well-rounded team and developing your plan for the business side. Establish a compelling, easy-to-navigate web presence and then identify a global list of investors to go after. Make sure not to underestimate the human resource commitment, and follow up often. Understanding the process, the time commitment and cost to execute a fund-raising campaign is half the battle. Being prepared and in context with the ever-changing cast of characters and the morphing investor landscape—most importantly, what investors are looking for—will allow you to have a far better chance of success.

ACKNOWLEDGMENTS
The authors would like to thank M. Quigley (VP, Market Research) and L. Parkinson (Director of Research), both of Life Science Nation, for their contributions in crafting this article.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

OPINION

# Pioneering government-sponsored drug repositioning collaborations: progress and learning

*Donald E. Frail, Madeleine Brady, K. Jane Escott, Alison Holt, Hitesh J. Sanganee, Menelas N. Pangalos, Chris Watkins and Craig D. Wegner*

Abstract | A new model for translational research and drug repositioning has recently been established based on three-way partnerships between public funders, the pharmaceutical industry and academic investigators. Through two pioneering initiatives — one involving the Medical Research Council in the United Kingdom and one involving the National Center for Advancing Translational Sciences of the National Institutes of Health in the United States — new investigations of highly characterized investigational compounds have been funded and are leading to the exploration of known mechanisms in new disease areas. This model has been extended beyond these first two initiatives. Here, we discuss the progress to date and the unique requirements and challenges for this model.

Drug repositioning — also referred to as repurposing, reprofiling, rescue, or indications discovery — is the process of identifying a new use for an existing drug or drug candidate in an indication outside the scope of the original indication. The case was made a decade ago that drug repositioning is one strategy to address the declining productivity of the pharmaceutical industry, as starting with a well-characterized compound decreases the duration of clinical development and could reduce attrition owing to issues such as poor pharmacokinetics or insufficient safety[1]. Furthermore, well-characterized clinical-stage compounds can be used to investigate novel disease hypotheses in human studies, which is important given the limitations of studies in animal models[2], including the lack of predictivity of the efficacy of compounds in subsequent clinical trials and issues with reproducibility (see the *Nature* focus on challenges in reproducible research). Making such compounds more widely available helps to drive hypothesis creation and can lead to broader testing in humans.

Multiple collaborations have since been established to enable the investigation of scientific advances within academia using drugs and drug candidates from industry. For example, in 2010, Pfizer's Indications Discovery Unit and the Washington University School of Medicine formed a partnership that provided access to proprietary data for a large portfolio of active and discontinued Pfizer drug candidates and allowed investigators to propose and collaborate on preclinical or clinical studies to investigate new uses[3,4]. Around the same time, the UK Medical Research Council (MRC) was seeking to enable a greater understanding of the mechanisms of human disease through experimental medicine studies enabled by access to high-quality compounds from industry. Although this strategic intent is distinct from that of a drug repositioning effort, the two complement each other. Meanwhile, Francis Collins, Director of the US National Institutes of Health (NIH), advocated for a more comprehensive repositioning programme that harnessed the strength of various stakeholders[5]. The NIH convened a roundtable of leading representatives from academia, the US government and private sector research and development (R&D) to explore such strategies, including one in which pharmaceutical companies would create a pool of compounds for further investigation by academia through a government grant programme[2]. Thus, in 2011 the MRC, in partnership with AstraZeneca, implemented the Mechanisms for Human Diseases Initiative, and in 2012 the NIH, through the National Center for Advancing Translational Sciences (NCATS), implemented the Discovering New Therapeutic Uses for Existing Molecules initiative.

The announcements of these programmes created some controversy, with some questioning the role of the NIH and NCATS as drug developers and the value of drug repositioning (for example, see REF. 6 and further discussion below). Given the timeframes of drug development, it may be several more years before the full value of these programmes is clear. Here, we make an interim assessment of the progress to date, discussing the unique requirements and challenges encountered with these programmes, as well as early indicators of success.

## Programme characteristics

***The MRC Mechanisms of Human Disease Initiative.*** This programme, a partnership between the MRC and AstraZeneca that was launched in 2011, provided academic researchers with unprecedented access to a high-quality collection of clinical and preclinical AstraZeneca compounds in order that they could propose new research into human disease mechanisms and the development of potential therapeutic interventions. The two partners had different but complementary motives. For the MRC, the initiative supported the MRC Translational Research Strategy, which has a strong emphasis on experimental medicine research to understand the biology of human disease and includes, where appropriate, preclinical studies using model systems. Successful preclinical studies could stimulate further pursuit of mechanistically related compounds in the clinic. The development of potential therapeutic interventions was not a primary goal, but it was hoped that successful studies and an increased understanding of the mechanisms involved in human disease, which would go into the public domain via peer-reviewed publications, would lead to the development of new medicines for patients by AstraZeneca or other companies. For AstraZeneca, the programme also provided the opportunity to build stronger relationships with members of the UK academic community with knowledge across a broad range of diseases.

Criteria were established for identifying which AstraZeneca compounds could be made available for external research. The focus was on identifying either development compounds that were discontinued and considered suitable for additional clinical studies (including confidence that target engagement is achievable), which were offered for clinical and preclinical proposals, or development compounds that were no

longer deemed suitable for clinical studies based on available clinical data (for example, lack of target coverage or sufficient safety) or limitations identified preclinically (for example, emerging preclinical long-term safety data), which were offered for preclinical proposals only.

AstraZeneca filtered through over 450 compounds that had been nominated for clinical development and identified a total of 22 suitable compounds (see Supplementary information S1 (table)). These compounds had been extensively characterized, including their potency, selectivity, pharmacology, complete pharmacokinetic and safety packages, target engagement and previous use in humans. For each compound, internal data were reviewed and summaries of the most relevant information were developed to enable investigators to craft a suitable proposal (see Supplementary information S2 (table)). This information was posted on the MRC website (that is, in the public domain) to attract the most innovative ideas from MRC-eligible investigators. At the time, this compendium of information was the largest single public source of what was proprietary information regarding efficacy, safety and other information on discontinued development compounds.

The MRC initiated a call for 'concept proposals', directing investigators to the compounds on the website. Over the 8-week open call period, more than 100 proposals were submitted from 37 different UK institutions, across all compounds and spanning a range of disease areas. The geographic diversity and breadth of ideas indicated that the concept of crowdsourcing was successful.

About half of the proposals fell within disease areas of focus for AstraZeneca, with the remainder falling into areas outside core therapeutic areas of interest, although whether the proposal fell within an area of interest to AstraZeneca was not a criterion used for review. A committee comprised of senior UK scientists convened by the MRC provided an initial review of the proposals for scientific merit, and independently, AstraZeneca provided a review of both the feasibility and suitability of the compound for the proposed investigations and the novelty of the studies. AstraZeneca was able to review these proposals owing to a confidentiality disclosure agreement (CDA) between AstraZeneca and the MRC, which covered these activities.

Informed by these reviews, a joint MRC–AstraZeneca committee then identified 25 proposals to advance to full proposal. The full proposals were collaboratively developed by the UK investigators and AstraZeneca scientists, and a champion of the proposal from AstraZeneca was required to proceed. Given the innovative funding approach taken, the collaboration between UK investigators and AstraZeneca scientists was essential for the development and pursuit of highly competitive and scientifically robust proposals. The full proposal development process was intense, with some complex clinical investigations being considered within short timelines. In all cases, AstraZeneca researchers were named co-applicants on the proposals, which were led by an academic. The full applications were considered by a specially convened MRC committee who were informed by international peer review comments; there was no AstraZeneca representation in the committee to ensure funding decisions were based solely on scientific merit and feasibility, without any commercial bias.

In total, 15 collaborative proposals — 7 preclinical and 8 clinical — were funded by the MRC and are now underway (TABLE 1). The collection of studies is highly diverse and each study is for a different indication, ranging from common disorders (for example, Alzheimer disease) to orphan diseases (for example, muscular dystrophy) and including indications outside of AstraZeneca's core areas of focus. Funding for the studies is provided to the investigators by the MRC (no MRC funds are provided to AstraZeneca). AstraZeneca, in addition to providing collaborative insight and suggestions to the full proposals, is responsible for providing the necessary drug supply and documents to support the regulatory and ethics committee filings by the investigators, as well as coordination of any adverse events when a compound is the subject of more than one study. The clinical studies are sponsored by the investigator, who is therefore ultimately responsible for the conduct of the study.

***NIH–NCATS Discovering New Therapeutic Uses for Existing Molecules.*** The pilot programme from NCATS was initiated in 2012 (see Further information). This programme matched NIH-funded researchers with a selection of 57 compounds previously discontinued from development (see Supplementary information S1 (table)). In this case, multiple companies were involved, with Abbott (supplying 3 compounds), Bristol-Myers Squibb (3), GlaxoSmithKline (4), Johnson & Johnson (3) and Sanofi (10) joining the founding companies of AstraZeneca (14), Eli Lilly (4) and Pfizer (17). Although similar in essence to the MRC programme, there were also substantial differences (TABLE 2):

- Purpose of the programmes. The primary goal of the MRC programme was to investigate mechanisms of human disease. Therefore, the MRC programme included preclinical studies, early concept-testing human studies and/or Phase II proof-of-concept studies. By contrast, the NCATS programme required the full proposal to contain a statistically powered Phase II proof-of-concept study. Neither the MRC programme nor the NIH programme was a global programme — each targeted only eligible investigators in the United Kingdom or the United States, respectively, although both were essential in building the foundation for a global programme as is discussed below (see also the AstraZeneca Open Innovation website).
- Compound criteria. The NCATS programme required compounds to have prior evidence of target coverage and manageable tolerability in humans, to potentially enable confident hypothesis testing in a new indication. The MRC programme also included compounds suitable for only preclinical use with evidence of potency, selectivity and exposure (typically by the oral route) in preclinical models. In both programmes, compounds were no longer in active development (that is, they were discontinued) to avoid any concern that public funding might be supplementing an ongoing commercial development objective.
- Review process. In the NCATS programme, the review of the concept proposals did not involve the companies, and proposals that were not selected for full proposal were not seen by the industry partners. The pharmaceutical companies were engaged under a CDA for those proposals selected for full development and could deny support of the proposal, thereby preventing full proposal submission. As with the MRC programme, the final funding decisions were made without company input.
- Template agreements. NCATS required each company to prepare and publicly post online template CDAs and collaborative research agreements (CRAs) at the time of the announcement of the programme, enabling early review by technology transfer offices and rapid implementation of such agreements. In addition, NCATS required a signed

Table 1 | **Government-sponsored collaborative drug repositioning projects**

| Project focus* | Collaborators, institution and industry partner | New indication (original indication) | Type of project |
|---|---|---|---|
| ***MRC-funded projects*** | | | |
| Saracatinib (AZD0530) as a novel analgesic for cancer-induced bone pain | • Dr D. Andrew, University of Sheffield, UK<br>• AstraZeneca | Cancer-induced bone pain (solid tumour) | Preclinical and clinical |
| Assessing the therapeutic efficacy of an 11βHSD1 inhibitor (AZD4017) in idiopathic intracranial hypertension | • Dr A. Sinclair, University of Birmingham, UK<br>• AstraZeneca | Idiopathic intracranial hypertension (diabetes and obesity) | Clinical |
| Evaluation of the selective endothelin A-receptor antagonist zibotentan (AZD4054) as a treatment for renal disease in systemic sclerosis (scleroderma) | • Professor C. Denton, University College London, UK<br>• AstraZeneca | Renal scleroderma (prostate cancer) | Clinical |
| Phase II study of the impact of a selective 11βHSD1 inhibitor (AZD4017) on biochemical markers of bone turnover in post-menopausal osteopaenia | • Professor P. Stewart, University of Leeds, UK<br>• AstraZeneca | Post-menopausal osteopaenia (diabetes and obesity) | Clinical |
| The role of $GABA_B$ receptor mechanisms in chronic cough (using AZD3355) | • Professor J. Smith, University of Manchester, UK<br>• AstraZeneca | Chronic cough (gastroesophageal reflux disease) | Clinical |
| Exploring $GABA_A$α2,3 signalling as novel therapy for peripheral neuropathies and primary dystonias (using AZD7325) | • Professor M. Koltzenburg, University College London, UK<br>• AstraZeneca | Dystonia or neuropathy (anxiety) | Clinical |
| SRC inhibitors (AZD0530) as potential antipsychotics: human testing with psilocybin | • Professor D. Nutt, Imperial College London, UK<br>• AstraZeneca | Psychosis (solid tumour) | Clinical |
| A new paradigm for testing pathway tractability in lung disease (using an MMP9 and MMP12 inhibitor (AZD1236)) | • Dr N. Hirani, University of Edinburgh, UK<br>• AstraZeneca | Idiopathic pulmonary fibrosis (chronic obstructive pulmonary disease) | Preclinical and clinical |
| The role of MMP inhibitors (AZD1236) in ameliorating muscular dystrophy | • Professor D. Wells, Royal Veterinary College, London, UK<br>• AstraZeneca | Muscular dystrophy (chronic obstructive pulmonary disease) | Preclinical |
| Investigating ATP regulation and P2X7 blockade (AZ11657312) in acute renal injury and its long-term complications | • Professor R. Unwin, University College London, UK<br>• AstraZeneca | Acute kidney injury (rheumatoid arthritis) | Preclinical |
| Efficacy of saracatinib (AZD0530) in treatment of chronic otitis media in preclinical mouse models | • Dr M. Cheeseman, University of Edinburgh, UK<br>• AstraZeneca | Chronic otitis media (solid tumour) | Preclinical |
| Evaluation of AZD1080 (GSK3β inhibitor) in a preclinical mouse model of motor neuron disease | • Dr R. Mead, University of Sheffield, UK<br>• AstraZeneca | Amyotrophic lateral sclerosis (Alzheimer disease) | Preclinical |
| Endothelin-1-mediated reduction of cerebral blood flow in Alzheimer disease: therapeutic potential of zibotentan (AZD4054) | • Professor S. Love, University of Bristol, UK<br>• AstraZeneca | Alzheimer disease (prostate cancer) | Preclinical |
| GSK3 as a multifunctional target for glioblastoma treatment; hitting multiple tumour hallmarks with a single drug (AZD2858) | • Dr S. Short, University of Leeds, UK<br>• AstraZeneca | Glioblastoma (Alzheimer disease) | Preclinical |
| ***NIH–NCATS-funded projects*** | | | |
| The efficacy and safety of a selective oestrogen receptor-β agonist (LY500307) | • Dr A. Breier, Indiana University, Indianapolis, USA<br>• Eli Lilly & Co. | Schizophrenia (benign prostatic hyperplasia) | Clinical |
| FYN inhibition by AZD0530 for Alzheimer disease | • Professor S. Strittmatter, Dr H. Nygaard and Professor C. Van Dyck, Yale University, New Haven, Connecticut, USA<br>• AstraZeneca | Alzheimer disease (solid tumour) | Clinical |
| Medication development of a novel therapeutic for smoking cessation | • Dr H. Brunzell, Virginia Commonwealth University, Richmond, Virginia, USA<br>• Dr K. Perkins, University of Pittsburgh, Pennsylvania, USA<br>• Janssen Research & Development, LLC | Smoking cessation (psoriasis and rheumatoid arthritis) | Clinical |

Table 1 (cont.) | **Government-sponsored collaborative drug repositioning projects**

| Project focus* | Collaborators, institution and industry partner | New indication (original indication) | Type of project |
|---|---|---|---|
| A novel compound for alcoholism treatment: a translational strategy | • Professor F. Akhlaghi, University of Rhode Island, Kingston, New York, USA<br>• Dr L. Leggio, National Institute on Alcohol Abuse and Alcoholism and National Institute on Drug Abuse, Bethesda, Maryland, USA<br>• Pfizer | Alcoholism (type 2 diabetes) | Clinical |
| Partnering to treat an orphan disease: Duchenne muscular dystrophy | • Dr K. Wagner, Kennedy Krieger Institute, Baltimore, Maryland, USA<br>• Dr S. Froehner, University of Washington, Seattle, USA<br>• Sanofi | Duchenne muscular dystrophy (not reported) | Clinical |
| Reuse of ZD4054 for patients with symptomatic peripheral artery disease | • Dr B. Annex, University of Virginia, Charlottesville, USA<br>• AstraZeneca | Peripheral arterial disease (prostate cancer) | Clinical |
| Therapeutic strategy for lymphangioleiomyomatosis (AZD0530 (saracatinib)) | • Dr T. Eissa, Baylor College of Medicine, Houston, Texas, USA<br>• AstraZeneca | LAM and TSC (solid tumour) | Clinical |
| Therapeutic strategy to slow progression of calcific aortic valve stenosis | • Dr J. Miller, Dr M. Enriquez-Sarano and Dr H. Schaff, Mayo Clinic, Rochester, New York, USA<br>• Sanofi | Calcific aortic valve stenosis (not reported) | Clinical |
| Translational neuroscience optimization of GlyT1 inhibitor | • Dr J. Krystal, Yale University, New Haven, Connecticut, USA<br>• Pfizer | Cognitive deficits in schizophrenia (schizophrenia) | Clinical |

*Compound codenames are provided where possible. 11βHSD1, 11β-hydroxysteroid dehydrogenase type 1; GABA, γ-aminobutyric acid; GlyT1, glycine transporter 1; GSK3, glycogen synthase kinase 3; LAM, lymphangioleiomyomatosis; MMP, matrix metalloproteinase; MRC, UK Medical Research Council; NCATS, National Center for Advancing Translational Sciences; NIH, National Institutes of Health; TSC, tuberous sclerosis complex.

CRA to be submitted with the final full proposal, which provided a deadline for any negotiations, whereas CRAs for the MRC programme were negotiated after the awards were granted to minimize the administrative burden of negotiating agreements with companies that might not have been awarded funding. The MRC process was somewhat facilitated by existing templates — the NIHR–MRC model Industry Collaborative Research Agreement (mICRA) and the UK Government Lambert agreements — which were further adapted for the purposes of this specific initiative. Despite the existence of the template agreements, some negotiation was still required.

Ultimately, nine clinical proposals were funded by the NIH, as summarized in TABLE 1 (see REF. 7 for an overview of this programme).

***National Research Program for Biopharmaceuticals.*** In 2013, a similar programme was established between AstraZeneca and the National Research Program for Biopharmaceuticals (NRPB) in Taiwan to facilitate translational research locally. This programme combines elements of the NIH–NCATS programme (for example, posting of a template CRA) and the MRC programme (both clinical and preclinical-only proposals supported). However, for the first time in these relationships, compounds actively being pursued in development ('live' compounds) at AstraZeneca were included, thereby setting a new precedent in this type of setting (see Supplementary information S1 (table)). One clinical and two preclinical projects were funded and are in progress. Additional proposals resulting from the networking and relationships established under this programme are now under collaborative discussions between the investigator, AstraZeneca and the NRPB.

**Responses to the programmes**

The announcement of the MRC programme was met with enthusiasm and seen as an exciting, unprecedented opportunity by UK investigators. When the NIH–NCATS programme was announced, it too was seen as an unprecedented opportunity, although there was some scepticism and criticism.

Some criticized the allocation of public funds to investigate company compounds and questioned the return to the public sector. However, academic clinical investigations, including those involving company compounds, are routinely supported by public funds, and these programmes provided access to compounds for mechanisms that have not previously been available. John LaMattina, former President of Research and Development at Pfizer, was one of many who criticized the involvement of academic investigators in drug development[6]. Such critics argued that the NIH (broadly used as a term to apply to academic researchers) lacks the experience required to develop drugs. Importantly, however, the intent of the programmes was to identify new uses of existing compounds and known mechanisms, not to develop drugs through to regulatory approval. Furthermore, previous experience of efforts to test company compounds in different tumour types, supported by the US National Cancer Institute (NCI), demonstrated that academic investigators are capable of achieving this goal. Examples in which NCI involvement in the early stages of development ultimately resulted in new medicines include cisplatin for the treatment of testicular, ovarian and lung cancer, and paclitaxel and fludarabine phosphate for the treatment of several cancers and lymphoma, respectively.
In addition, projects supported by the

NCATS programme required company collaboration to develop the final proposal, thus combining the experience and knowledge of both academic and industry investigators.

Another criticism was that company scientists have already exhaustively considered alternative indications for their compounds. This is not always true, as companies typically have therapeutic areas of focus and only invest time and money in those. The indications considered by these programmes were unlimited and utilized the broad expertise lying outside pharmaceutical companies. More importantly, science continuously progresses with new discoveries, and some of these discoveries could lead to a new use for an existing compound, for example, unpublished data from investigators linking a target to a disease. Furthermore, the totality of the available data may not be sufficient for investment by the company, and additional preclinical or clinical data may change interest levels.

There was also criticism of these programmes because the compound structures were not initially disclosed[8]. This criticism primarily came from academic groups doing *in silico* drug repositioning, in which structural features of one compound are found to be similar to those of a drug with a known effect that acts through a different target (off-target activity). However, the purpose of the programmes was hypothesis-based repositioning with defined mechanisms of action, not to generate new hypotheses for an individual molecule working through off-target biology or other repositioning approaches. It was concluded that non-hypothesis-based investigations were not within the scope of these programmes and that the extensive compound information provided to investigators was sufficient to meet the goals of the programme, thus compound structures were not required. Nevertheless, Southan *et al.* did take the initiative to search publicly available databases to identify most, if not all, of the compound structures in the NIH–NCATS programme, although the accuracy was not confirmed by the companies (see REF. 8 and the Southan figshare page).

Caution and scepticism was also evident within pharmaceutical companies. Many clinical compounds, although listed as discontinued after failing to improve on the current standard of care in their initial Phase II indication, remain under preclinical evaluation for alternative indications. This status typically persists for a few years, during which time R&D leaders can be reluctant to allow the compound to be tested externally. By the time this internal deliberation is exhausted, interest in the compound and its mechanism of action, both within and outside the company, and the patent life have diminished. In addition, as R&D budgets have decreased over the past decade many companies have narrowed their therapeutic area focus, leading to hesitation by some to support investigator-sponsored trials in non-core therapy areas. Budget constraints can therefore make it challenging for companies to supply the clinically formulated drug (and matched placebo) substance, updated regulatory documents (for example, the Investigator's Brochure and Chemical, Manufacturing and Controls section) and scientific and clinical compound-specific advice for disease areas that are not a high priority. Recent activities, described below, indicate that more and more companies have overcome these initial concerns.

**Unique requirements and challenges**

Clinical trials generally fall into one of two groups: company-sponsored studies or investigator-sponsored studies. To date, investigator-sponsored studies, regardless of the funder, almost always use live compounds that are in development or on the market. In these circumstances, existing project teams provide the required compound insight, regulatory document updates, safety database access and so on, and they are ultimately the decision makers for whether a given study should be run. For a discontinued compound, given that the project team has often been disbanded, there are unique challenges that require new ways of working to be established.

***Compound selection.*** The initial MRC and NIH programmes were constructed to include discontinued compounds as these provided the easiest initial path to develop and implement such a groundbreaking activity. However, the status of a compound can be dynamic. As an example, one discontinued compound, a hormone modulator, was repositioned within the company for polycystic ovarian syndrome and a Phase II study was initiated while the MRC programme was being established. In this case, one MRC proposal was similar to the internal programme and led to a separate collaboration with a leading investigator in the area that focused on key scientific questions for the programme. It is therefore important in such collaborations to be flexible and enable a compound to be used in more than one collaboration or to be re-evaluated for development within the company.

The following general criteria were established by AstraZeneca and agreed to by the MRC and NCATS to select compounds for these publicly sponsored collaborative drug-repositioning programmes:

- Clinical and/or preclinical evidence of potency, selectivity and exposure supporting target coverage was required to ensure conclusive testing of a novel mechanism-driven hypothesis.
- For clinical proof-of-concept testing, sufficient patient safety to support further development was required. This involved substantial analysis and judgment in the context of the indication, length of study or target patient population. The length of supporting safety studies was highlighted to alert investigators of the acceptable length of proposed clinical studies. Investigators could include longer-term toxicology studies in their proposals if necessary.
- Reasonable cross-species activity and suitability for dosing in animals was required for all compounds made available for non-clinical studies.
- Patent life was not required. However, sufficient remaining patent life, new intellectual property or regulatory data exclusivity would probably be required to advance positive findings further.
- There had to be no other relevant commitments or complex legal agreements (for example, a compound is not partnered with or licensed to a third party).
- For compounds to be used for clinical studies, enough drug substance to support a reasonably sized clinical trial had to be available. The same route of administration as previously used was required to avoid time, cost and attrition risk associated with new formulation development, safety studies and human pharmacokinetic studies. Doses were typically limited to those supported by the existing data, though in rare cases additional preclinical data were obtained as part of the programme. The existing clinical data were directed to new patient populations and requirements (for example, inclusion of women of child bearing potential), which was challenged by regulatory authorities at times.
- Approval for use by the company project team (if appropriate) and the therapy area head was required.

When selecting compounds, one of the largest challenges faced was the collation and generation of the relevant datasets, particularly when project teams had been disbanded.

Table 2 | **Key differences between the MRC and NIH–NCATS programmes***

| Characteristic | MRC | NIH–NCATS |
|---|---|---|
| Purpose | Understanding the mechanisms of disease | Therapeutic development |
| Scope | Preclinical or clinical or both | Translational projects to hypotheses tested in full Phase IIa trials |
| Proposal review | AstraZeneca is involved in reviewing the concept proposals | The NIH performs peer-review of concept proposals without company involvement |
| Industry participants | AstraZeneca only | Eight pharmaceutical companies (AstraZeneca, Pfizer, Eli Lilly & Co., Abbott, Bristol-Myers Squibb, GlaxoSmithKline, Johnson & Johnson and Sanofi) |
| Time from call for proposals to dosing the first subject in the first study | 25 months (sequential process) | 14 months (more parallel activities) |
| Collaborative agreement | Negotiated after MRC project approval using a mICRA template | A signed agreement is required before the final proposal submission. Company template CRA agreements posted online |
| Funding | 3-year award provided up front | Milestone-driven, first year provided up front |
| Milestone management | Extendable | Preclinical milestone is ≤12 months after funding is awarded; total time allocated ≤36 months |
| Supervision | No required collaboration meetings | NCATS convenes regular investigator–company collaboration meetings |

*There are also several similarities between the two initiatives, including the use of template agreements and of discontinued compounds only. Companies provide the compounds, regulatory documents and advice, and the MRC or NIH provides funding to the investigators; no funding goes to the pharmaceutical companies. CRA, collaborative research agreement; mICRA, model Industry Collaborative Research Agreement; MRC, Medical Research Council; NCATS, National Center for Advancing Translational Sciences; NIH, National Institutes of Health.

For example, final analysis of histopathology data from a rodent carcinogenicity study of a discontinued project resulted in a new reportable finding for one compound. In another example, one discontinued programme was found to still be on partial clinical hold under the original investigational new drug (IND) application, requiring notification to the UK Investigator. In many cases, Investigator's Brochures had not been updated and needed to be. One of the key learnings for AstraZeneca has been that every project should be fully closed out, including noting incomplete datasets, with a view to potential project repositioning or project reactivation. Final project document summaries should be quickly completed by the original team.

The AstraZeneca partnership with the NRPB in Taiwan and the more recent NIH–NCATS 'Round 2' programmes allowed the inclusion of live compounds. This makes the collation of project data much simpler as there is a project team in place, although the dataset is also more actively evolving as ongoing studies read out. In addition, live compounds bring their own challenges in terms of balancing confidentiality, focus and intellectual property rights to new inventions. Often quoted is the concern that externally sponsored research (be it clinical or non-clinical) may generate data with negative implications for the original programme (for example, a new safety signal). This is best addressed by engagement of the project team and vetting of the risks versus benefits (risk–benefit assessment) of the additional indication to the entire portfolio, not just the single project, to avoid withholding more novel and interesting live compounds.

***Funding and programme management.*** Funding for the studies was provided directly from the public funding body to the principal investigator. It should be noted that the 'in-kind' costs incurred by the companies are not trivial, particularly if remanufacture of the active product, preparation of drug product and placebo, completion of study reports or regulatory documents, and/or patient safety coordination are needed.

The processes and timelines for the implementation of the projects varied between the MRC and the NIH–NCATS programmes, which resulted in differences in timelines to the start of studies, particularly for clinical programmes (TABLE 2). This can lead to issues such as expiration of the clinical drug supply. Differences between the programmes included:

- The NIH–NCATS programme required the inclusion of signed collaboration agreements with the full proposal submission, eliminating the post-funding delay caused by negotiating such agreements, whereas there was no such deadline for negotiated agreements for the MRC collaborations. A balance is needed between the speed of processing applications and reducing unnecessary paperwork associated with ultimately unsuccessful applications.
- For the MRC projects, delays were encountered between the approval of the grant and the institution receiving the funds. For example, some investigators needed to hire personnel for the work and the institution would not allow recruiting to begin until funds were received.
- The MRC committed the full funding at the beginning of the grant and was flexible in allowing extensions of the grant timeline in some circumstances, whereas the NIH–NCATS programme was milestone-driven, with the first year of funding being provided up front and subsequent funding being granted based on the study achieving certain milestones.
- The process in the United Kingdom for implementing certain clinical studies was more sequential (that is, various ethics, regulatory and study-implementation approvals were obtained in sequence), whereas investigators in the United States approached certain activities in parallel.
- NIH–NCATS required some supervision and regular meetings to assess the progress of studies towards milestones.

Nevertheless, the first clinical study results for the MRC and AstraZeneca NIH–NCATS programmes were obtained in the first half of 2014; one example is a mechanistic study from the MRC programme that evaluated the effects of a $GABA_B$ receptor agonist, AZD3355, on capsaicin-induced cough in healthy volunteers, and another, from the NIH–NCATS programme, is a Phase Ib safety study of saracatinib in patients with Alzheimer disease. Both projects progressed to the second phase of their proposals. Most clinical study results from the MRC and NIH–NCATS projects will be obtained in 2015–2016.

*Collaborative working.* There tend to be strict rules governing externally sponsored studies to ensure the avoidance of influence or bias by the pharmaceutical company over the independent investigator. Concerns regarding possible conflicts of interest are particularly acute for marketed products or compounds in Phase III development. These repositioning programmes did not include such late-stage compounds. Indeed, collaboration was essential to bring together the best scientific and clinical expertise to support these approaches, which were aimed at testing novel hypotheses. This led to the creation of a new AstraZeneca policy for collaborative discussions regarding the construction of full proposals for these investigator-sponsored studies. Beyond the scientific, disease and compound properties, others areas of discussion included the trial design, statistical power, patient safety and decision criteria. Nevertheless, because these remain investigator-sponsored studies, the ultimate responsibility and decisions for all aspects of the study lie with the investigator.

*Pharmacovigilance.* One key element of these collaborative studies was to ensure the appropriate ownership and management of patient safety (pharmacovigilance). The responsibility for industry-sponsored studies rests with the company, who maintain one global Investigator's Brochure and a global safety database to support and ensure harmonization with regulatory and ethics safety-reporting requirements. For externally sponsored studies, the investigator takes on the risk–benefit assessment for the study and the regulatory reporting, with additional requirements to report back to the parent company regarding safety.

Regulatory agencies clearly devolve responsibility for individual studies to the investigator, and this raised the potential for lack of centralized coordination for discontinued compounds — a concern if one compound spawned multiple studies in different territories (for example, raising the question of how to effectively share an observation of a suspected unexpected serious adverse reaction (SUSAR) related to the compound in a US-based study with a UK-based investigator working with the same compound in a separate study). Therefore, AstraZeneca developed a system whereby the company retains ownership of the Investigator's Brochure and required annual updates from investigators to keep this as one core document. Study-specific risk–benefit assessments are documented through a cover note to the Investigator Brochure and/or the protocol and/or the regulatory filing documents (investigational medicinal product dossier (IMPD) or IND documents). The patient safety database is owned by the company and contracted to a third party for maintenance, timely reporting from investigators and dissemination to all sites working with the same compound. Furthermore, safety data owned by the various investigators are made accessible to other investigators. Thus, AstraZeneca provides a central position in the cross-study evaluation of the safety profile and maintains one harmonized and globally available Investigator's Brochure.

*Intellectual property and publications.* Frequently asked questions regarding these partnerships typically involved intellectual property (IP) and publication rights. In general, any existing IP remained with the owning party and options were established for the company to license any new IP generated by the investigator. Independent of IP, a company may wish to license the data generated by the investigator to support further development. For successful studies not further progressed by the compound originator, it is in the best interest of the company and the academic researchers to find a way to advance the programme for the benefit of patients. Regarding publications, the investigators retained the right to publish with standard provisions that provide for the company to review before submission. Both potential issues were avoided through early discussions regarding motivations of each partner when designing the programmes.

## Early indicators of success

Although these novel models have only been running for a short time in the context of drug development, there are early indicators of success and benefits. Highlights include:

- Greater sharing of valuable, and previously closely guarded, proprietary information: both the open publication of compound information, providing a single summary source of previously unavailable information, and the template legal agreements are unprecedented.
- Crowdsourcing works: there was a diverse range of proposals across institutions for indications beyond those initially anticipated that were submitted within compressed timeframes.
- Improved quality of grants: the collaborative discussions during the construction of full proposals brought together academia and industry to produce higher quality proposals, as suggested by the assessments given by the peer review committees and the 100% success rate of obtaining ethics and regulatory approvals for studies.
- Novel translational research that may not otherwise have been pursued: companies were not pursuing these studies and these studies would not have been possible for investigators to conduct without access to the proprietary compounds and public funds.
- Generation of new intellectual property: to date, at least two patents have been filed by investigators as a direct result of these activities.
- Two clinical studies have been completed, enabling progression to the next phase of the MRC- or NIH-funded projects.
- Spin-off proposals: proposals made by investigators but not funded by the programme still found a way of moving forward. In one case, AstraZeneca elected to fund a proposal on its own merits, and in another case the investigator accessed alternative sources of funding.
- There has also been the very important benefit of forging closer engagement between the academic and commercial sectors, reducing the perceived barriers and misperceptions, and expanding the knowledge base of compound information.

Saracatinib (also known as AZD0530), a deprioritized compound in the AstraZeneca portfolio, is an especially interesting case study for these repositioning programmes. It is a potent, orally bioavailable inhibitor of SRC tyrosine kinase family members, including FYN kinase, with an expansive preclinical and clinical foundation of research, including Phase II studies in a variety of solid tumours. Ultimately, saracatinib failed to sufficiently modify disease progression in these trials. The general and reproductive toxicology of saracatinib was studied in rat and dog models for up to 6 months. In human studies, the safety profile was such that further clinical investigation was possible. Although the focus had been in oncology, it is now recognized that the SRC kinase family is involved in biology across multiple organ systems. Between the MRC and NCATS programmes, five preclinical or clinical projects, and one additional spin-off project, were funded for further investigation across a diverse range of non-oncological diseases (BOX 1). Data were recently published for one project, showing that saracatinib was effective in reducing spatial memory defects and synaptic depletion in a mouse model of Alzheimer disease[9].

These results support the evaluation of saracatinib in patients with Alzheimer disease, and clinical studies are taking place as part of this programme[10].

Perhaps the best indicator of early success is the expansion of these programmes. As previously mentioned, AstraZeneca and the NRPB in Taiwan partnered to initiate a similar program, and both the MRC and the NIH have initiated a second round of calls for proposals and funding. The MRC programme has expanded beyond AstraZeneca to include 68 compounds from 7 pharmaceutical companies (AstraZeneca, GlaxoSmithKline, Johnson & Johnson, Eli Lilly & Co., Pfizer, Takeda and UCB). The NIH programme has expanded to include live development compounds and support for paediatric proposals.

Ultimately, the true measure of success will be either positive clinical data that define a new mechanism involved in impacting human disease or definitive negative data that help to disprove a hypothesis and result in a redirection of attention and resources to more promising avenues.

**Box 1 | The evolving science of saracatinib**

Saracatinib (also known as AZD0530) is an especially interesting case study for these repositioning programmes. Saracatinib inhibits the SRC kinase family, and although it was developed for oncology indications, it is now recognized that SRC kinases could be important in diseases related to multiple organ systems. Saracatinib is being investigated in six non-oncological indications as a result of the UK Medical Research Council (MRC) and US National Institutes of Health (NIH) programmes (see the figure).

**SRC kinase as a novel analgesic for cancer-induced bone pain**

The hypothesis that SRC kinase is a crucial component of cancer-induced bone pain will be tested through preclinical studies in an animal model of bone cancer pain, investigating the effects of inhibiting SRC on pain-related behaviour, spinal cord neuron phosphorylation and signalling, as well as bone resorption, to identify potential analgesic mechanisms of SRC inhibition. A randomized controlled trial of saracatinib will investigate whether it has analgesic effects in cancer patients with bone metastases.

**SRC inhibitors as potential antipsychotics: human testing with psilocybin**

Existing data demonstrate that the SRC kinase pathway is directly involved in the observable symptoms associated with the acute administration of hallucinogens that modulate the 5-$HT_{2A}$ receptor, such as psilocybin. This clinical study is evaluating the role of SRC kinase in blocking psychosis induced by infusion of psilocybin.

**Efficacy of saracatinib in treatment of chronic otitis media**

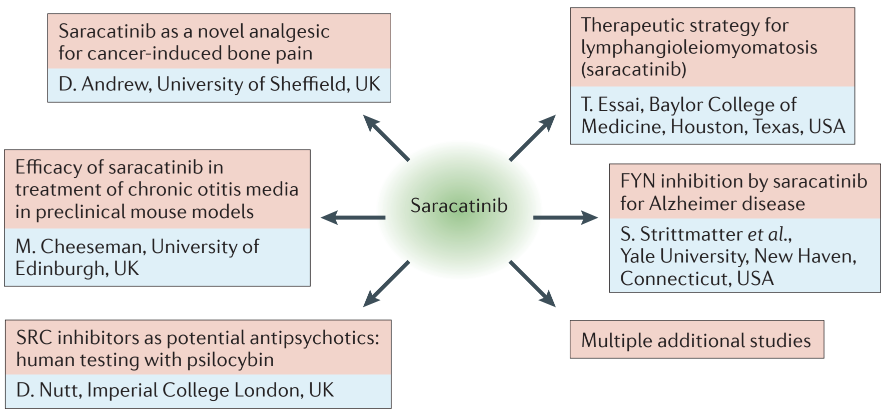
This preclinical study will determine whether local SRC kinase inhibition moderates vascular leak and bulla fluid accumulation leading to reduced hearing loss in models of chronic otitis media.

**Therapeutic strategy for lymphangioleiomyomatosis and tuberous sclerosis**

Lymphangioleiomyomatosis (LAM) is a rare progressive cystic lung disease. This research team discovered that SRC kinase is active in LAM cells and is important for cell growth and a cell's ability to move around and invade tissues. This preclinical study aims to determine whether blocking SRC activity is safe and can reduce the growth and the spread of LAM cells.

**FYN inhibition by saracatinib for Alzheimer disease**

FYN, a SRC kinase family member, is implicated in triggering Alzheimer disease. This study seeks to test the hypothesis that FYN has an important role in Alzheimer disease and that saracatinib provides benefit to patients with Alzheimer disease in a Phase IIa clinical study. This effort has shown that saracatinib has beneficial effects in a mouse model of Alzheimer disease[9], and a Phase Ib study of the safety and tolerability of saracatinib in patients with Alzheimer disease has been completed[10]. The drug is to be tested for effectiveness in slowing disease progression in a larger population with mild Alzheimer disease.




## Outlook

These pioneering government-sponsored drug repositioning initiatives have successfully piloted a new era of open collaboration and innovation between academia and the pharmaceutical industry on translational research. However, these initiatives have only begun to scratch the surface of the potential opportunities to strengthen and expand such efforts.

The number of compounds involved could easily increase by the participation of additional companies and incorporation of more live compounds, particularly those in early development, as is often done in the field of oncology. Additionally, consideration of off-target effects could be included. Disease-centric charity organizations that have focused their strategies on the translation of research into therapies for patients could partner to fund programmes of interest on compounds that have been made publicly available. Some pioneering charities have embraced repositioning as a strategy, including the Michael J. Fox Foundation, which funded their first repositioning-focused programme in 2010, the Leukaemia and Lymphoma Society and Cancer Research UK (CRUK), although the opportunity remains to specifically leverage the compounds posted publicly.

Start-up companies could contribute compounds to expand their efforts without financial dilution. Once a company has participated in one partnership the effort needed for additional partnerships is much lower. Venture capitalists could establish new start-up or spin-out companies based on either the compounds made available or the funded proposals. In fact, one could imagine a 'marketplace' of compounds and funders for investigators to access. As a result of the success of current pilot programmes, AstraZeneca has recently launched a broader worldwide 'open innovation' initiative that offers a range of compounds for which academic investigators can submit new repositioning ideas and translational research (see Further information). This broad open innovation platform is the first to offer compounds for collaboration and provide a template for a true marketplace that invites ideas and proposals from any contributor and any sector.

A more radical model would use crowdsourcing during the creation of the proposals. At present, proposals are submitted by individuals. Alternatively, an initial proposal, submitted in an open manner, could then be crowdsourced for improvements, including alternative patient populations, endpoints, or trial designs. 'Gamification' — the use of game thinking and mechanics to engage users in solving problems — is one potential crowdsourcing route wherein funders, investigators, companies and reviewers each play a different part. Another possibility is the platform developed by Transparency Life Sciences to develop drugs through collaborative input. Any crowdsourcing model would have to resolve intellectual property rights, determining who performs the research and how credit for the final proposal is assigned.

Barriers remain that could prevent this new partnership model from reaching its full potential. Creating broad awareness of the opportunity among investigators can be challenging. The availability of new compounds should increase over time; however, the available clinical supply of discontinued compounds quickly becomes exhausted. Resistance to the inclusion of live compounds remains prevalent within some companies and funders (although the most recent NIH–NCATS programme did support the inclusion), and the inclusion of biologics is challenged by the complexities of compound supply and delivery. Patent lives are continuously diminishing and the pursuit of more expensive registration studies following positive early studies may be deterred by the lack of financial return. Although regulatory data exclusivity provisions may provide an alternative to patent life, the limited term available in the United States is likely to be insufficient in many cases.

The traditional barriers to open innovation have only just begun to be addressed. Supplemental resourcing, especially funding, is still needed to justify the pursuit of certain areas, including those involving rare and niche indications or those compounds with a limited remaining patent life. If projects are successful, the appropriate distribution of royalties must recognize the relative contributions of each party. Even greater trust in sharing between parties can be achieved, including sharing scientific information, patient safety monitoring and scientific credit for innovative advances.

Nevertheless, the achievements of these pioneering initiatives include important lessons learned for all parties involved and advances that tackle many of the key barriers. The question confronting all parties now is: will we move forward against any remaining hurdles and advance this model or will we revert to the historical more 'closed' translational research and collaboration models?

*Donald E. Frail was previously at the Emerging Innovations Unit, Scientific Partnering & Alliances, AstraZeneca, 35 Gatehouse Drive, Waltham, Massachusetts 02451, USA.*
*Present address: Research and External Science and Innovation, Allergan, Inc., 2525 Dupont Drive, Irvine, California 92612, USA.*

*Madeline Brady was previously at the Emerging Innovations Unit, Scientific Partnering & Alliances, AstraZeneca, Melbourn Science Park, Cambridge Road, Melbourn, Herts SG8 6EE, UK.*
*Present address: Global Medical Affairs Oncology, Global Medicines Development, Innovative Medicines and Early Development Biotech Unit, AstraZeneca, Melbourn Science Park, Cambridge Road, Melbourn, Herts SG8 6EE, UK.*

*K. Jane Escott, Alison Holt, and Hitesh J. Sanganee are at the Emerging Innovations Unit, Scientific Partnering & Alliances, Innovative Medicines and Early Development Biotech Unit, AstraZeneca, Melbourn Science Park, Cambridge Road, Melbourn, Herts SG8 6EE, UK*

*Menelas Pangalos is at the Innovative Medicines and Early Development Biotech Unit, AstraZeneca, Melbourn Science Park, Cambridge Road, Melbourn, Herts SG8 6EE, UK.*

*Chris Watkins is at the Medical Research Council, 14th Floor, One Kemble Street, London WC2B 4AN, UK.*

*Craig D. Wegner is at the Emerging Innovations Unit, Scientific Partnering & Alliances, AstraZeneca, 35 Gatehouse Drive, Waltham, Massachusetts 02451, USA.*

*Correspondence to D.E.F.*
*e-mail: frail_don@allergan.com*

doi:10.1038/nrd4707
Published online 20 November 2015

1. Ashburn, T. T. & Thor, K. B. Drug repositioning: Identifying and developing new uses for existing drugs. *Nat. Rev. Drug Discov.* **3**, 673–683 (2004).
2. Cook, D. *et al.* Lessons learned from the fate of AstraZeneca's drug pipeline: a five-dimensional framework. *Nat. Rev. Drug Discov.* **13**, 419–431 (2014).
3. Mullard, A. Could pharma open its drug freezers? *Nat. Rev. Drug Discov.* **10**, 399–400 (2011).
4. Carroll, J. Washington U gets second look at Pfizer molecules. *Fierce Biotech* [online], http://www.fiercebiotech.com/story/washington-u-researchers-get-second-look-pfizer-molecules/2010-05-18#ixzz33uYuYIr8 (2010).
5. Collins, F. S. Mining for therapeutic gold. *Nat. Rev. Drug Discov.* **10**, 397 (2011).
6. LaMattina, J. The NIH is going to discover drugs… Really? *Forbes* [online], http://www.forbes.com/sites/johnlamattina/2012/05/15/the-nih-is-going-to-discover-drugs-really/ (2012).
7. Colvis, C. M. & Austin, C. P. The NIH-industry New Therapeutic Uses pilot program: demonstrating the power of crowdsourcing. *Drug Repurp. Rescue Repos.* **1**, 15–16 (2015).
8. Southan, C., Williams, A. J. & Ekins, S. Challenges and recommendations for obtaining chemical structures of industry-provided repurposing candidates. *Drug Discov. Today* **18**, 58–70 (2013).
9. Kaufman, A. C. *et al.* Fyn inhibition rescues established memory and synapse loss in Alzheimer mice. *Ann. Neurol.* **77**, 953–971 (2015).
10. Strittmatter, S. M. Safety and tolerability of AZD0530 (Saracatinib) in Alzheimer's disease. *ClinicalTrials.gov* [online], https://clinicaltrials.gov/ct2/show/NCT01864655?term=strittmatter&rank=1 (2014).

### Acknowledgements
The UK Medical Research Council (MRC) and the US National Institute of Health (NIH) initiatives were supported by a large range of individuals within the MRC and the NIH as well as external investigators, and investigators from AstraZeneca and Medimmune. At the risk of excluding many, in particular the authors would like to thank J. Latimer (MRC), C. Colvis and B. Dunn (NIH–NCATS), G. Wilkinson, C. Wilks, A. Longton, S. Curran, K. Hickling and the members of the New Opportunities Emerging Innovations Unit (AstraZeneca). The authors would like to acknowledge the contribution made by all those involved and the exciting ideas and proposals received from academic investigators.

### Competing interests statement
The authors declare competing interests: see Web version for details.

### FURTHER INFORMATION
**AstraZeneca Open Innovation:** http://openinnovation.astrazeneca.com/
**MRC Translational Research Strategy:** http://www.mrc.ac.uk/research/initiatives/experimental-medicine
***Nature* focus on challenges in reproducible research:** http://www.nature.com/nature/focus/reproducibility/
**NCATS Round 2 Industry provided agents (2014):** http://www.ncats.nih.gov/research/reengineering/rescue-repurpose/therapeutic-uses/directory2014.html
**NCATS Pilot Programme (2012):** http://www.ncats.nih.gov/research/reengineering/rescue-repurpose/therapeutic-uses/directory.html
**NIHR/MRC model Industry Collaborative Research Agreements:** http://www.nocri.nihr.ac.uk/resources/micra/
**NIH–NCATS: Discovering New Therapeutic Uses for Existing Molecules. 2012 URL:** http://www.ncats.nih.gov/research/reengineering/rescue-repurpose/therapeutic-uses/therapeutic-uses.html
**Southan figshare page:** http://figshare.com/articles/NCats_Compounds_with_identifications/92850
**MRC-Industry Asset Sharing Initiative (2014):** http://www.mrc.ac.uk/news-events/news/world-s-largest-collection-of-deprioritised-pharma-compounds-opens-to-researchers/
**UK Government Lambert agreements:** https://www.gov.uk/model-agreements-for-collaborative-research

### SUPPLEMENTARY INFORMATION
See online article: S1 (table) | S2 (table)

**ALL LINKS ARE ACTIVE IN THE ONLINE PDF**

# Why we need risk innovation

If emerging technologies such as nanotechnology are to reach their full potential we need to radically change our approach to risk, argues **Andrew D. Maynard**.

In October 2014, Google announced it was working on an innovative nanotechnology-based approach to avoiding and managing disease[1]. The idea was to create a pill that would deliver magnetic, functionalized nanoparticles from the gut to the bloodstream. Once there, they would circulate — presumably for days, or longer — picking up biomarkers of disease along the way. The particles would then be remotely interrogated directly by the patient, perhaps using a wrist-mounted monitor. In effect, the plan was to create the ultimate in wearable tech: a personal device that could give you up-to-the-minute information on health and wellness, much as wrist-worn devices provide feedback on fitness today.

Google's nanosensor concept is certainly audacious. Its success though will depend on overcoming a number of challenges — not least, addressing potential risks. Based on what is currently known about nanoparticle behaviour, the technology faces a plethora of possible health and environmental challenges. Failure to address these could leave the company with a non-starter on its hands. Yet the probability of causing harm is not the only risk that could prevent these nanosensors from becoming a reality. In the expanded list of potential risks, there is also the chance of outmoded or overly restrictive regulations blocking progress; or the possibility of investor ambivalence, consumer suspicion, or social media backlash. These hint at a much larger and murkier risk landscape that emerging technologies will have to navigate to be successful.

## An emerging risk landscape

Google's nanoparticle sensors are indicative of a growing number of technologies that are facing increasingly complex risk-related challenges. Recently, the Future of Life Institute awarded close to US$7 million for research aimed at ensuring the robust and beneficial development of artificial intelligence[2] — funding prompted by how unexpected risks could undermine the technology's development[3]. Earlier this year, published research into using the gene-editing technique CRISPR/Cas9 on human embryos sparked an international discussion around the ethics and safety of such techniques[4]. And as self-driving cars move towards becoming a reality on public roads, debate around potential risks is intensifying[5].

These and many more emerging technologies face an uncertain future because of a growing disconnect between the rate at which we are innovating, and our ability to assess and manage the adverse consequences of this innovation. And this is not simply a problem of minimizing risks to human health and the environment. Important as evidence-based health and environmental risk assessment and management are, they fail to capture the full panoply of personal, social, environmental, technological, economic, political and corporate risks that determine the fate of new technologies. All play a significant and growing role in determining the success or failure of emerging products and capabilities, and together form a complex and interconnected risk landscape that cannot be navigated without a similarly complex and multidimensional understanding of risk.

## 'Aba-made'

In the Nigerian city of Aba, entrepreneurs are becoming adept at reverse-engineering and repurposing products for the local market. It's a phenomenon known as 'Aba-made', and is synonymous with the city's informal economic vitality and entrepreneurialism. While Aba-made is predominantly associated with recreating designer shoes, bags and clothing, there are signs that the city's artisans are becoming more sophisticated in the use of new technologies[6]. How far Aba entrepreneurs will extend their technological skills isn't yet clear. Yet as technology innovation and entrepreneurship continue to blossom in sub-Saharan Africa[7], Aba-made applications of emerging technologies certainly present a plausible future, and one that is likely to flourish with little formal risk oversight.

This plausible 'Aba-made' future echoes a growing trend in technology innovation democratization around the world, stimulated by an ever-lower entry barrier to using cutting-edge technologies. It's a trend that potentially allows local needs and opportunities to be responded to, precisely because it exists at the fringes of formal regulatory frameworks. Yet it also raises the spectre of unanticipated risks and unintended consequences. And in today's interconnected world, local adverse impacts can have a profound influence on the technology's global development.

In the US and beyond, 'do-it-yourself' science and technology is similarly shaking up the risk landscape. The Maker Movement, for instance, is leading a revolution in opening up individual and community access to sophisticated technologies[8]. Similarly, community labs are increasingly enabling individuals to play around — quite literally — with technologies such as synthetic biology. These movements operate largely outside the confines of established organizations and oversight frameworks. But despite often slipping through the regulatory net, they are rarely risk-agnostic. Far from it — there is often a community ethic that takes the consequences of actions seriously[9]. For instance, DIYBio.org — a self-identified 'Institution for Do-It-Yourself Biologists' — encourages members to ask a panel of professional biosafety experts about their questions on safety and risk. Yet these movements are changing perspectives on risk in ways that potentially destabilize an already fragile network of formal regulations and policies. And at the heart of this disruption are shifts in what is considered to be of value, and how it is potentially threatened by risk. Within this moving risk landscape, human health and environmental security remain critically important. But they are joined by a long list of additional factors that include, but are not limited to, social justice, community resilience, fiscal independence, and personal discovery and pleasure. The result is a broadening out of what constitutes risk, and a need for innovation in how to successfully navigate an evolving risk landscape.

## The democratization of influence

At the same time, the evolution of this landscape is being stimulated by increasing global democratization of influence and information. Social media, and the internet more broadly, have all but eliminated

geographical, national and cultural barriers to organization, advocacy and influence. Citizens from different countries and cultures now have the capacity to band together within virtual constituencies, and influence action on risks far from their physical location. Increasingly, this influence is fuelled by perceptions, beliefs and values that are not always grounded in scientific evidence, yet nevertheless have societal legitimacy.

To complicate matters further, even evidence-based approaches to assessing, managing and regulating risk are often grounded in values, with community norms guiding how risk is defined and evaluated. For example, the current European definition of nanomaterials for regulatory purposes, which helps frame the identification of risks and subsequent responses to them, reflects a belief in what is important and implementable, not necessarily what has the potential to cause harm[10,11].

## Risk innovation

If we are to succeed in building value through emerging technologies such as nanotechnology, we need a radical new approach to risk — one that matches and complements the inventiveness and transformative nature of technology innovation, and provides the means to navigate successfully through an evolving risk landscape. We need, in effect, parallel innovation in how we conceptualize risk and use this knowledge to good effect — we need, I would argue, a new domain of research and practice: risk innovation.

Risk innovation can be thought of as an organizing framework for generating new understanding, insights and inventions around risk; and translating these into products, tools and practices that protect social and environmental value, as well as enabling its creation and growth. It's an approach that has the potential to generate radical new insights into navigating the risk landscape. As a framework, it gives license to what might be described as risk entrepreneurship, where the ultimate measure of an idea's worth is whether it has an impact, not whether it adheres to convention. And as in technology entrepreneurship, it encourages a culture of experimentation — a culture grounded in transdisciplinarity, creativity and imagination; and epitomized by serendipity and a 'fail fast fail forward' mentality that recognizes the importance of failure in developing robust solutions to both challenges and opportunities.

As a concept, risk innovation frames risk as a threat to existing or future 'value', where value is broadly and multiply defined within personal, societal and organizational contexts. This in turn supports a definition of innovation, from the perspective of risk, as a process of generating new knowledge, ideas, and inventions, and translating these into concepts, products or processes that protect this value.

## Risk innovation in practice

In 2014, I was involved in organizing a workshop with the Dutch design organization V2_ Institute for the Unstable Media on exploring the nature and meaning of 'responsible innovation'. Participants represented a broad range of disciplines, including engineering, business, medicine, art and design, and language. The outcome was a book of seventeen haiku, combining poems, abstract images and expositions around responsibility in innovation — an unusual result from an academic meeting[12]. The book was designed to capture the nuances of our insights in a way that an academic paper could not. But it was also aimed at stimulating new ideas in its readers that would lead to further innovation in responsible technology development among entrepreneurs and innovators. As an output, it has more in common with arts and literature than it does risk analysis. And yet just as these modes of communication and engagement reveal novel perspectives, the book potentially opens up risk navigation pathways to its readers that would otherwise remain hidden.

Such artistic collaborations define one end of the risk innovation spectrum. At the other end lie initiatives that are almost exclusively grounded in science and technology. In 2008 for instance, the US Environmental Protection Agency and the National Institutes of Health launched an ambitious new programme to transform toxicology testing[13]. The Toxicology in the 21st Century programme is designed to use emerging techniques in high-throughput screening, computational biology and data processing to evaluate the potential risks of tens of thousands of untested chemicals. It's an innovative initiative that is already leading to changes in how potential risks associated with chemicals are identified and addressed, and is paving the way towards the safer, more effective use of substances in consumer and commercial products.

These two very different examples illustrate how creative collaborations and novel application within the framework of risk innovation can shake up conventional thinking in ways that opens up new possibilities. Yet they only scratch the surface of what is possible. Risk innovation has the potential to reveal new pathways through complex risk landscapes. It encourages a sophisticated dialogue around building and maintaining value in a world where risk is not only endemic, but integral to progress. It complements approaches to ensuring the safe and responsible use of emerging technologies such as responsible innovation[14] and anticipatory governance[15]. And it has the potential to open up routes to addressing risk that would otherwise be closed — whether these are technological, social, economic or political.

Without risk innovation, all we are left with is business as usual. And for technologies such as Google's nanoparticle sensors, this isn't likely to be good news. ❐

*Andrew D. Maynard directs the Risk Innovation Lab at Arizona State University, PO Box 875603, ASU, Tempe, Arizona 85387-5603, USA.*
*e-mail: andrew.maynard@asu.edu*

## References

1. Drahl, C. *Chem. Eng. News* **92,** 26–28 (2014).
2. *New International Grants Program Jump-Starts Research to Ensure AI Remains Beneficial* (Future of Life Institute, 2015); http://futureoflife.org/AI/2015selection
3. *Research Priorities for Robust and Beneficial Artificial Intelligence* (Future of Life Institute, 2015).
4. Reardon, S. *Nature* http://dx.doi.org/10.1038/nature.2015.17410 (2015).
5. Lerner, P. The hurdles facing autonomous vehicles. *Automobile* (22 June 2015); http://go.nature.com/6JXgTz
6. Effiong, U. Nigerians will soon have to worry about implanted pacemaker security. That's fantastic. *Slate* (2015); http://go.nature.com/UQstrM
7. Bright, J. & Hruby, A. The rise of silicon savannah and Africa's tech movement. *TechCrunch* (2015); http://go.nature.com/23YYLG
8. Bajarin, T. Why the Maker Movement is important to America's future. *Time* (19 May 2014); http://go.nature.com/JDmtVK
9. Grushkin, D., Kuiken, T. & Millet, P. *Seven Myths & Realities About Do-It-Yourself Biology* (Woodrow Wilson Int. Center for Scholars, 2013).
10. Maynard, A. D. *Nature* **475,** 31 (2011).
11. European Commission *Towards a Review of the EC Recommendation for a Definition of the Term "Nanomaterial": Part 2: Scientific-technical Evaluation of Options to Clarify the Definition and to Facilitate its Implementation* (Publications Office of the European Union, 2015).
12. Maynard, A. D. *Responsible Innovation – Seventeen Haiku* (2014); http://2020science.org/2014/12/23/responsible-innovation-seventeen-haiku/
13. Hartung, T. *Nature* **460,** 208–212 (2009).
14. Stilgoe, J., Owen, R. & Macnaghten, P. D *Res. Policy* **42,** 1568–1580 (2013).
15. Guston, D. H. *Social Stud. Sci.* **44,** 218–242 (2014).